

The Information Retrieval Series

Nicola Ferro  
Carol Peters *Editors*

# Information Retrieval Evaluation in a Changing World

Lessons Learned from 20 Years of CLEF



Springer

*Series Editors*

ChengXiang Zhai

Maarten de Rijke

*Editorial Board*

Nicholas J. Belkin

Charles Clarke

Diane Kelly

Fabrizio Sebastiani

More information about this series at <http://www.springer.com/series/6128>


Nicola Ferro • Carol Peters  
Editors

# Information Retrieval Evaluation in a Changing World

Lessons Learned from 20 Years of CLEF

 Springer

*Editors*

Nicola Ferro   
Dipartimento di Ingegneria  
dell'Informazione  
Università degli Studi di Padova  
Padova, Italy

Carol Peters  
Consiglio Nazionale delle Ricerche  
Istituto di Scienza e Tecnologie  
dell'Informazione  
Pisa, Italy

ISSN 1387-5264

The Information Retrieval Series

ISBN 978-3-030-22947-4

ISBN 978-3-030-22948-1 (eBook)

<https://doi.org/10.1007/978-3-030-22948-1>

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG.  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Foreword

Search is ubiquitous today: finding facts, people, how-to instructions, images, maps, etc. are all services taken for granted. But this has not always been the case, and going back even 20 years ago would find information seeking much more difficult, requiring manual location of appropriate printed material or appropriate friendly experts.

A primary scientific discipline behind the success of today's search algorithms is information retrieval, where research has been ongoing since the mid-1950s. And one of the major driving forces of this discipline has been a strong requirement for solid evaluation of experimental results, beginning with the Cranfield experiments in the early 1960s.

An extension of the Cranfield tradition was the TREC (Text REtrieval Conference) which started in 1992 to evaluate search algorithms for retrieval of English text. In 2000, the cross-language evaluation for European languages moved from TREC to form the CLEF conference (Cross-Language Evaluation Forum) where it gained new life, both with improved test collections and enthusiastic participation.

This book chronicles the evolution of CLEF in the last 20 years from the initial cross-language evaluation start to its incredibly multi-faceted version today. Its impact goes far beyond the improved algorithms for cross-language retrieval.

CLEF continued monolingual and cross-language evaluation for more than 9 years, expanding from 3 languages to over 21 languages, including ones from outside of Europe. The impact of these multilingual evaluations was threefold: the test collections built for these languages, the resulting improved search algorithms and, most importantly, the opportunity for increasing numbers of (mostly) academic organizations across Europe to get deeply involved with information retrieval and evaluation. The ability to work in native languages was one attraction, but another was the opportunity to be part of a large evaluation effort within Europe.

The co-operative effort required to assemble test collections across different languages has led CLEF to have a loose confederation, where the organizations initially involved in creating multilingual text collections then started proposing new tasks, along with providing the evaluation for these tasks. Evaluation of image retrieval started in 2003 using multilingual annotations, along with evaluation of

multilingual speech retrieval. These tasks not only attracted new communities of researchers but were mostly tightly tied to specific practical domain problems.

The transition of CLEF to the Conference and Labs of the Evaluation Forum in 2010 expanded even further the broadening of the evaluation scope. Not only did tracks like ImageCLEF morph into labs working with medical images, identification of plants from images or birds from audio soundtracks, but specific domains were targeted with new evaluation labs, such as work with patents and various health applications or the use of XML to improve search involving metadata. Constant throughout this expansion has been the emphasis on solid evaluation practice and on the natural multilingual aspects of these tasks.

CLEF has also been heavily involved in furthering retrieval evaluation itself, including infrastructure systems like DIRECT to appropriately store data and results, and new paradigms of evaluation such as living labs with real users and tasks.

A summary of the impact of CLEF has many facets. First, there is the impact of improved search, including multilingual text search, image search, etc. that have been detailed in various publications over the 20 years. Second, there is the increased involvement of many European groups in evaluation—a look at the table of contents of this book provides ample proof of this. And third, there is the conference itself, with its emphasis on solid evaluation of research aimed at real-world problems.

National Institute of Standards and Technology  
Gaithersburg, MD, USA  
April 2019

Donna Harman

# Preface

CLEF—the *Cross-Language Evaluation Forum* for the first 10 years, and the *Conference and Labs of the Evaluation Forum* since—is an international initiative whose main mission is to promote research, innovation and development of information retrieval (IR) systems through the organization of annual experimental evaluation campaigns.<sup>1</sup> The aim of this volume is to celebrate the 20th anniversary of CLEF and to trace its evolution over these first two decades, as it has kept pace with and often anticipated current trends in information management, with the results helping to stimulate progress in the field of IR system experimentation and evaluation.

In order to do this, the volume is divided into six parts. Parts **I** and **II** provide background and context. The first three chapters in Part **I** explain what is intended by experimental evaluation and the underlying theory, describing how this has been interpreted in CLEF and in other internationally recognized evaluation initiatives. In addition, the introductory chapter illustrates the activity and results of CLEF over the years in some detail. Part **II** presents research architectures and infrastructures that have been developed to manage experimental data and to provide evaluation services in CLEF and elsewhere. Parts **III**, **IV** and **V** represent the core of the volume, consisting of a series of chapters presenting some of the most significant evaluation activities in CLEF, ranging from the early multilingual text processing exercises to the later, more sophisticated experiments on multimodal collections in diverse genre and media. In all cases, the focus has not only been on describing “what has been achieved” but most of all on “what has been learnt”. The final part is dedicated to examining the impact CLEF has had on the research world and to discussing current and future challenges, both academic and industrial. In particular, the concluding chapter discusses the relevance of IR benchmarking in an industrial setting. Clearly, the ultimate aim of an activity of this type must be the involvement of real-world user communities. IR research can never be considered only at the theoretical level; the over-riding factors are the requirements of society at large.

---

<sup>1</sup>CLEF in French means “key”, which gave us our symbolic logo and the somewhat cryptic pronunciation (kle).



Before ending this brief preface, it is incumbent on us to acknowledge our gratitude to all the people, groups and institutions who have collaborated with us in the organization and day-to-day running of CLEF, almost always on a purely voluntary basis. Unfortunately, it is impossible to mention them one by one. There are just too many, it would take pages and pages, and then we would probably have inadvertently forgotten someone. More than a few appear as authors of chapters in this volume—but over the course of the years, many more have been involved. So here, we limit ourselves to naming just a very few people without whom this endeavour would have been inconceivable.

As stated in the first chapter, CLEF began life in 1997 as a track for *Cross-Language Information Retrieval* within the Text REtrieval Conference (TREC) series, hosted by the US National Institute of Standards and Technology (NIST). After 3 years of experiments, Donna Harman of NIST and Peter Schäuble of Eurospider Information Technology AG, Zurich, agreed that this activity would be better located in the multilingual environment of Europe. The very first CLEF campaign was thus organized in 2000, in a collaboration between NIST, represented by Donna Harman and Ellen Voorhees; the Institute of Information Science and Technologies, CNR, Pisa; and Eurospider, Zurich. CLEF 2000 culminated in a workshop at the European Conference on Digital Libraries, held that year in Lisbon, at which the first results and strategy for future editions were discussed. Thus, in the early years, the technical coordination of CLEF was centred in Zurich, at Eurospider, under the leadership of Martin Braschler, now at the Zurich University of Applied Sciences, and the scientific coordination at CNR, Pisa, with the support of Costantino Thanos, leader of DELOS, Network of Excellence on Digital Libraries, funded by the European Commission. Michael Kluck, Informationszentrum Sozialwissenschaften (IZ), Bonn/Berlin, was also very much involved. Several years later, the research group led by Maristella Agosti, University of Padua, entered the core coordinating team, taking over responsibility for the management and processing of the test collections and developing tools to handle them and later on also becoming the scientific coordinators. We are enormously grateful to Donna, Ellen, Peter, Martin, Michael, Costantino and Maristella, for their initial support and backing. Without their early efforts, CLEF would not exist today. In addition, we remember all the help and advice we have received throughout the years from our various Steering Committees. CLEF is truly a joint endeavour, so many people have contributed to its success.

In addition we really must thank our editorial board, who gave us so much advice, especially during the early days, in the preparation of this project. Our appreciation also goes to the long list of reviewers who, with their painstaking work of comments and suggestions to the authors of the various chapters in this volume,

greatly helped to improve the quality of the contents. And finally, our gratitude goes to the editorial team at Springer led by Ralf Gerstner and to Chengxiang Zhai and Maarten de Rijke, editors of Springer's Information Retrieval Series, whose interest and encouragement enabled us to bring this work to fruition.

Padua, Italy  
Pisa, Italy  
April 2019

Nicola Ferro  
Carol Peters

# Contents

## Part I Experimental Evaluation and CLEF

|  |    |
|--|----|
| <b>From Multilingual to Multimodal: The Evolution of CLEF over Two Decades</b> ..... | 3  |
| Nicola Ferro and Carol Peters  |    |
| <b>The Evolution of Cranfield</b> .....  | 45 |
| Ellen M. Voorhees  |    |
| <b>How to Run an Evaluation Task</b> .....   | 71 |
| Tetsuya Sakai  |    |

## Part II Evaluation Infrastructures

|  |     |
|--|-----|
| <b>An Innovative Approach to Data Management and Curation of Experimental Data Generated Through IR Test Collections</b> ..... | 105 |
| Maristella Agosti, Giorgio Maria Di Nunzio, Nicola Ferro, and Gianmaria Silvello   |     |
| <b>TIRA Integrated Research Architecture</b> .....   | 123 |
| Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein   |     |
| <b>EaaS: Evaluation-as-a-Service and Experiences from the VISCERAL Project</b> .....   | 161 |
| Henning Müller and Allan Hanbury   |     |

## Part III Multilingual and Multimedia Information Retrieval

|  |     |
|--|-----|
| <b>Lessons Learnt from Experiments on the Ad Hoc Multilingual Test Collections at CLEF</b> ..... | 177 |
| Jacques Savoy and Martin Braschler   |     |
| <b>The Challenges of Language Variation in Information Access</b> .....                          | 201 |
| Jussi Karlgren, Turid Hedlund, Kalervo Järvelin, Heikki Keskustalo, and Kimmo Kettunen           |     |

|  |     |
|--|-----|
| <b>Multi-Lingual Retrieval of Pictures in ImageCLEF</b> .....  | 217 |
| Paul Clough and Theodora Tsikrika  |     |
| <b>Experiences from the ImageCLEF Medical Retrieval and Annotation Tasks</b> .....   | 231 |
| Henning Müller, Jayashree Kalpathy-Cramer, and Alba García Seco de Herrera   |     |
| <b>Automatic Image Annotation at ImageCLEF</b> .....   | 251 |
| Josiah Wang, Andrew Gilbert, Bart Thomee, and Mauricio Villegas  |     |
| <b>Image Retrieval Evaluation in Specific Domains</b> .....  | 275 |
| Luca Piras, Barbara Caputo, Duc-Tien Dang-Nguyen, Michael Riegler, and Pål Halvorsen   |     |
| <b>About Sound and Vision: CLEF Beyond Text Retrieval Tasks</b> .....  | 307 |
| Gareth J. F. Jones   |     |
| <b>Part IV Retrieval in New Domains</b>  |     |
| <b>The Scholarly Impact and Strategic Intent of CLEF eHealth Labs from 2012 to 2017</b> .....  | 333 |
| Hanna Suominen, Liadh Kelly, and Lorraine Goeruiot   |     |
| <b>Multilingual Patent Text Retrieval Evaluation: CLEF-IP</b> .....  | 365 |
| Florina Piroi and Allan Hanbury  |     |
| <b>Biodiversity Information Retrieval Through Large Scale Content-Based Identification: A Long-Term Evaluation</b> .....   | 389 |
| Alexis Joly, Hervé Goëau, Hervé Glotin, Concetto Spampinato, Pierre Bonnet, Willem-Pier Vellinga, Jean-Christophe Lombardo, Robert Planqué, Simone Palazzo, and Henning Müller |     |
| <b>From XML Retrieval to Semantic Search and Beyond</b> .....  | 415 |
| Jaap Kamps, Marijn Koolen, Shlomo Geva, Ralf Schenkel, Eric SanJuan, and Toine Bogers  |     |
| <b>Part V Beyond Retrieval</b>   |     |
| <b>Results and Lessons of the Question Answering Track at CLEF</b> .....   | 441 |
| Anselmo Peñas, Álvaro Rodrigo, Bernardo Magnini, Pamela Forner, Eduard Hovy, Richard Sutcliffe, and Danilo Giampiccolo   |     |
| <b>Evolution of the PAN Lab on Digital Text Forensics</b> .....  | 461 |
| Paolo Rosso, Martin Potthast, Benno Stein, Efstathios Stamatatos, Francisco Rangel, and Walter Daelemans   |     |
| <b>RepLab: An Evaluation Campaign for Online Monitoring Systems</b> .....  | 487 |
| Jorge Carrillo-de-Albornoz, Julio Gonzalo, and Enrique Amigó   |     |

**Continuous Evaluation of Large-Scale Information Access Systems:  
A Case for Living Labs** ..... 511  
Frank Hopfgartner, Krisztian Balog, Andreas Lommatzsch, Liadh Kelly,  
Benjamin Kille, Anne Schuth, and Martha Larson

**Part VI Impact and Future Challenges**

**The Scholarly Impact of CLEF 2010–2017** ..... 547  
Birger Larsen

**Reproducibility and Validity in CLEF** ..... 555  
Norbert Fuhr

**Visual Analytics and IR Experimental Evaluation** ..... 565  
Nicola Ferro and Giuseppe Santucci

**Adopting Systematic Evaluation Benchmarks in Operational Settings** .... 583  
Jussi Karlgren

**Author Index** ..... 591

**Subject Index** ..... 593

# Acronyms

|          |   |     |
|----------|---|-----|
| ACLIA    | Advanced CrossLingual Information Access . . . . .                                      | 84  |
| ADL      | Activities of Daily Living . . . . .  | 294 |
| ANOVA    | ANalysis Of VAriance . . . . .  | 577 |
| AP       | Average Precision . . . . .   | 571 |
| ASR      | Automatic Speech Recognition . . . . .  | 308 |
| BIIRRR   | Barriers to Interactive IR Resources Re-use . . . . .                                   | 429 |
| CHiC     | Cultural Heritage in CLEF . . . . .   | 422 |
| CHIIR    | Conference on Human Information Interaction<br>and Retrieval . . . . .                  | 429 |
| CIRCO    | Coordinated Information Retrieval Components<br>Orchestration . . . . .                 | 19  |
| CLAIRE   | Combinatorial visual Analytics system for Information<br>Retrieval Evaluation . . . . . | 577 |
| CLEF     | Conference and Labs of the Evaluation Forum . . . . .                                   | 586 |
| CLIR     | Cross-Language Information Retrieval . . . . .  | 307 |
| CL-SDR   | Cross-Language Spoken Document Retrieval . . . . .                                      | 308 |
| CL-SR    | Cross-Language Speech Retrieval . . . . .   | 314 |
| CNN      | Convolutional Neural Network . . . . .  | 291 |
| CNN-LSTM | CNN Long Short-Term Memory Network . . . . .  | 292 |
| COLD     | COsy Localization Database . . . . .  | 278 |
| CUI      | Concept Unique Identifier . . . . .   | 336 |
| DAS      | Discriminative Accumulation Scheme . . . . .  | 280 |
| DCG      | Discounted Cumulated Gain . . . . .   | 572 |
| DIRECT   | Distributed Information Retrieval Evaluation Campaign<br>Tool . . . . .                 | 105 |
| DL       | Digital Library . . . . .   | 115 |
| DM       | Data Mining . . . . .   | 565 |
| eHealth  | Electronic Health . . . . .   | 335 |
| EHR      | Electronic Health Record . . . . .  | 334 |
| ERR      | Expected Reciprocal Rank . . . . .  | 95  |
| ERR-IA   | Intent-Aware Expected Reciprocal Rank . . . . .   | 96  |

|          |   |     |
|----------|---|-----|
| FIRE     | Forum for Information Retrieval Evaluation . . . . .                    | 46  |
| GoP      | Grid of Points . . . . .  | 577 |
| HTTP     | HyperText Transfer Protocol . . . . .                                   | 113 |
| IAP      | Interpolated Average Precision . . . . .                                | 289 |
| ICD      | International Classification of Diseases . . . . .                      | 340 |
| IE       | Information Extraction . . . . .  | 334 |
| INEX     | INitiative for the Evaluation of XML Retrieval . . . . .                | 417 |
| IoU      | Intersection over Union . . . . .                                       | 290 |
| IR       | Information Retrieval . . . . .   | 565 |
| iSBS     | Interactive Social Book Search . . . . .                                | 422 |
| IV       | Information Visualization . . . . .                                     | 565 |
| Khresmoi | Knowledge Helper for Medical and Other Information<br>users . . . . .   | 350 |
| LD       | Linked Data . . . . .   | 420 |
| LOD      | Linked Open Data . . . . .  | 105 |
| LRT      | Lifelog Retrieval Task . . . . .  | 294 |
| LST      | Lifelog Summarization Task . . . . .                                    | 294 |
| MAP      | Mean Average Precision . . . . .  | 312 |
| MC2      | Microblog Contextualization . . . . .                                   | 429 |
| MF       | Mean F-measure . . . . .  | 290 |
| MIAP     | Mean Interpolated Average Precision . . . . .                           | 290 |
| MIMIC    | Multiparameter Intelligent Monitoring in Intensive Care . . . . .       | 338 |
| MIR      | Multilingual Information Retrieval . . . . .                            | 307 |
| MLIA     | MultiLingual Information Access . . . . .                               | 19  |
| NCU      | Normalised Cumulative Utility . . . . .                                 | 95  |
| nDCG     | normalized Discounted Cumulative Gain . . . . .                         | 575 |
| nERR     | normalised Expected Reciprocal Rank . . . . .                           | 96  |
| NLP      | Natural Language Processing . . . . .                                   | 334 |
| NTCIR    | NII Testbeds and Community for Information access<br>Research . . . . . | 323 |
| OCR      | Optical Character Recognition . . . . .                                 | 209 |
| OOV      | Out Of Vocabulary . . . . .   | 212 |
| ORP      | Open Relevance Project . . . . .  | 118 |
| PAAPL    | Passive-Aggressive with Averaged Pairwise Loss . . . . .                | 291 |
| PIR      | Personalized Information Retrieval . . . . .                            | 33  |
| QPA      | Query Performance Analyzer . . . . .                                    | 579 |
| qrels    | query-relevance set . . . . .   | 93  |
| RBP      | Rank-Biased Precision . . . . .   | 575 |
| RDF      | Resource Description Framework . . . . .                                | 106 |
| RF       | Relevance Feedback . . . . .  | 420 |
| RR       | Reciprocal Rank . . . . .   | 95  |
| RTHSD    | Randomised Tukey Honestly Significant Difference . . . . .              | 97  |
| SBS      | Social Book Search . . . . .  | 420 |
| SCR      | Spoken Content Retrieval . . . . .                                      | 316 |
| SCST     | Supporting Complex Search Tasks . . . . .                               | 427 |

|                   |  |     |
|-------------------|--|-----|
| SDR               | Spoken Document Retrieval . . . . .                                | 312 |
| SERP              | Search Engine Result Page . . . . .                                | 79  |
| ShARe             | Shared Annotated Resources . . . . .                               | 338 |
| SME               | Statistics–Metrics-Experiments . . . . .                           | 569 |
| SNOMED CT         | Systematized Nomenclature of Medicine—Clinical Terms . . . . .     | 334 |
| SR                | Snippet Retrieval . . . . .  | 420 |
| STC               | Short Text Conversation . . . . .                                  | 84  |
| SVM               | Support Vector Machines . . . . .                                  | 280 |
| TAR               | Technology Assisted Reviews . . . . .                              | 336 |
| TC                | Tweet Contextualization . . . . .                                  | 420 |
| TME               | Topics–Metrics-Experiments . . . . .                               | 567 |
| TREC              | Text REtrieval Conference . . . . .                                | 421 |
| TRECVID           | TREC Video Retrieval Evaluation . . . . .                          | 313 |
| UMLS              | Unified Medical Language System . . . . .                          | 334 |
| VA                | Visual Analytics . . . . .   | 565 |
| VAIRĚ             | Visual Analytics for Information Retrieval Evaluation . . . . .    | 570 |
| VATE <sup>2</sup> | Visual Analytics Tool for Experimental Evaluation . . . . .        | 573 |
| VDM               | Visual Data Mining . . . . .                                       | 565 |
| VIRTUE            | Visual Information Retrieval Tool for Upfront Evaluation . . . . . | 571 |
| WHO               | World Health Organization . . . . .                                | 334 |
| WSD               | Word Sense Disambiguation . . . . .                                | 13  |
| WWW               | We Want Web . . . . .  | 72  |
| XML               | eXtensible Markup Language . . . . .                               | 294 |



# Editorial Board

Martin Braschler, Zurich University of Applied Sciences (ZHAW), Switzerland

Paul Clough, University of Sheffield, UK

Julio Gonzalo, National Distance Education University (UNED), Spain

Donna Harman, National Institute of Standards and Technology (NIST), USA

Gareth Jones, Dublin City University, Ireland

Jussi Karlgren, Gavagai and KTH Royal Institute of Technology, Sweden

Henning Müller, University of Applied Sciences Western Switzerland (HES-SO), Switzerland

Maarten de Rijke, University of Amsterdam, The Netherlands

Paolo Rosso, Universitat Politècnica de València, Spain

Jacques Savoy, University of Neuchâtel, Switzerland

# Reviewers

Giuseppe Amato, ISTI, National Council of Research (CNR), Italy

Sameer Antani, National Library of Medicine, National Inst. of Health (NIH), USA

Leif Azzopardi, University of Strathclyde, UK

Ben Carterette, Spotify and University of Delaware, USA

Paul Clough, University of Sheffield, UK

Giorgio Maria Di Nunzio, University of Padua, Italy

Fabrizio Falchi, ISTI, National Research Council (CNR), Italy

Norbert Fuhr, University of Duisburg-Essen, Germany

Costantino Grana, University of Modena and Reggio Emilia, Italy

Michael Granitzer, University of Passau, Germany

Gregory Grefenstette, Institute for Human & Machine Cognition (IHMC) and Biggerpan, USA

Donna Harman, National Institute of Standards and Technology (NIST), USA

Bogdan Ionescu, University Politehnica of Bucharest, Romania

Noriko Kando, National Institute of Informatics (NII), Japan

Makoto P. Kato, Kyoto University, Japan

Aldo Lipani, University College London, UK

David Losada, University of Santiago de Compostela, Spain

Bernardo Magnini, Fondazione Bruno Kessler (FBK), Italy

Maria Maistro, University of Copenhagen, Denmark

Miguel Martinez, Signal Media, UK

Piero Molino, Uber AI Labs, USA

Manuel Montes-y-Gómez, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico

Henning Müller, University of Applied Sciences Western Switzerland (HES-SO), Switzerland

Douglas Oard, University of Maryland, USA

Martin Potthast, Leipzig University, Germany

Andreas Rauber, Vienna University of Technology, Austria

Kirk Roberts, The University of Texas at Dallas, USA

Paolo Rosso, Universitat Politècnica de València, Spain

Tony Russell-Rose, UXLabs, UK

Michail Salampanis, Alexander Technological Educational Institute (ATEI) of Thessaloniki, Greece

Mark Sanderson, RMIT University, Australia

Jacques Savoy, University of Neuchâtel, Switzerland

Giuseppe Serra, University of Udine, Italy

Gianmaria Silvello, University of Padua, Italy

Assaf Spanier, The Hebrew University of Jerusalem, Israel

Suzan Verberne, Leiden University, The Netherlands

Christa Womser-Hacker, University of Hildesheim, Germany

**Part I**  
**Experimental Evaluation and CLEF**

# From Multilingual to Multimodal: The Evolution of CLEF over Two Decades



Nicola Ferro and Carol Peters

**Abstract** This introductory chapter begins by explaining briefly what is intended by experimental evaluation in information retrieval in order to provide the necessary background for the rest of this volume. The major international evaluation initiatives that have adopted and implemented in various ways this common framework are then presented and their relationship to CLEF indicated. The second part of the chapter details how the experimental evaluation paradigm has been implemented in CLEF by providing a brief overview of the main activities and results obtained over the last two decades. The aim has been to build a strong multidisciplinary research community and to create a sustainable technical framework that would not simply support but would also empower both research and development and evaluation activities, while meeting and at times anticipating the demands of a rapidly evolving information society.

## 1 Introduction

CLEF—the Cross-Language Evaluation Forum for the first 10 years, and the Conference and Labs of the Evaluation Forum since—is an international initiative whose main mission is to promote research, innovation, and development of information retrieval systems.

---

N. Ferro (✉)

Department of Information Engineering, University of Padua, Padova, Italy  
e-mail: [ferro@dei.unipd.it](mailto:ferro@dei.unipd.it)

C. Peters

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (ISTI), National Research Council (CNR), Pisa, Italy  
e-mail: [carol.peters@isti.cnr.it](mailto:carol.peters@isti.cnr.it)

© Springer Nature Switzerland AG 2019

N. Ferro, C. Peters (eds.), *Information Retrieval Evaluation in a Changing World*, The Information Retrieval Series 41,  
[https://doi.org/10.1007/978-3-030-22948-1\\_1](https://doi.org/10.1007/978-3-030-22948-1_1)

CLEF currently promotes research and development by providing an infrastructure for:

- multilingual and multimodal system testing, tuning and evaluation;
- investigation of the use of unstructured, semi-structured, highly-structured, and semantically enriched data in information access;
- creation of reusable test collections for benchmarking;
- exploration of new evaluation methodologies and innovative ways of using experimental data;
- discussion of results, comparison of approaches, exchange of ideas, and transfer of knowledge.

This activity is conducted by providing a platform for experimental system evaluation and then holding workshops and organizing an annual conference where researchers and developers can get together to discuss results and exchange ideas and experiences.

This aim of this chapter is to present the activity and results of CLEF over the last two decades. In Sect. 1, we begin by explaining briefly what is intended by experimental evaluation in information retrieval, providing pointers to more detailed discussions, in particular to the other two chapters in this first part of the book, in order to provide the necessary context. We then present the major international evaluation initiatives that have adopted this common framework and indicate their relationship to CLEF.

Sections 2 and 3 detail how the experimental evaluation paradigm has been implemented in CLEF by providing a brief overview of the main activities and results obtained in these first 20 years. The evolution and shift in focus can be seen as a reflection of the development of the information retrieval scene in this span of time. While the activities of CLEF in the first 10 years (2000–2009) were very much focused on the evaluation of systems developed to run on multiple languages, since 2010 the scope has been widened to embrace many different types of multimodal retrieval. For convenience, in this chapter we refer to these two distinct, but not separate, phases of CLEF as CLEF 1.0 and CLEF 2.0. Figure 1 shows clearly the evolution of CLEF over the last two decades, and the shift from mainly text retrieval in the early years of CLEF 1.0 to all kinds of multimedia retrieval, with increasing attention being given to dynamic and user-oriented activities in CLEF 2.0. Many of the main CLEF activities are described in separate chapters in Parts III, IV and V of this volume; however, full details on all experiments, including methodologies adopted, test collections employed, evaluation measures used and results obtained, can be found in the CLEF Working Notes<sup>1</sup> and the CLEF Proceedings.<sup>2</sup>

Section 4 provides valuable information on the test collections that have been created as a result of the evaluation activities in CLEF and on their availability. The final two Sections describe the CLEF Association, established in 2013 to support

---

<sup>1</sup>Published annually in the CEUR Workshop Proceedings series ([CEUR-WS.org](http://www.ceur-ws.org)).

<sup>2</sup>Published by Springer in their Lecture Notes for Computer Science series.

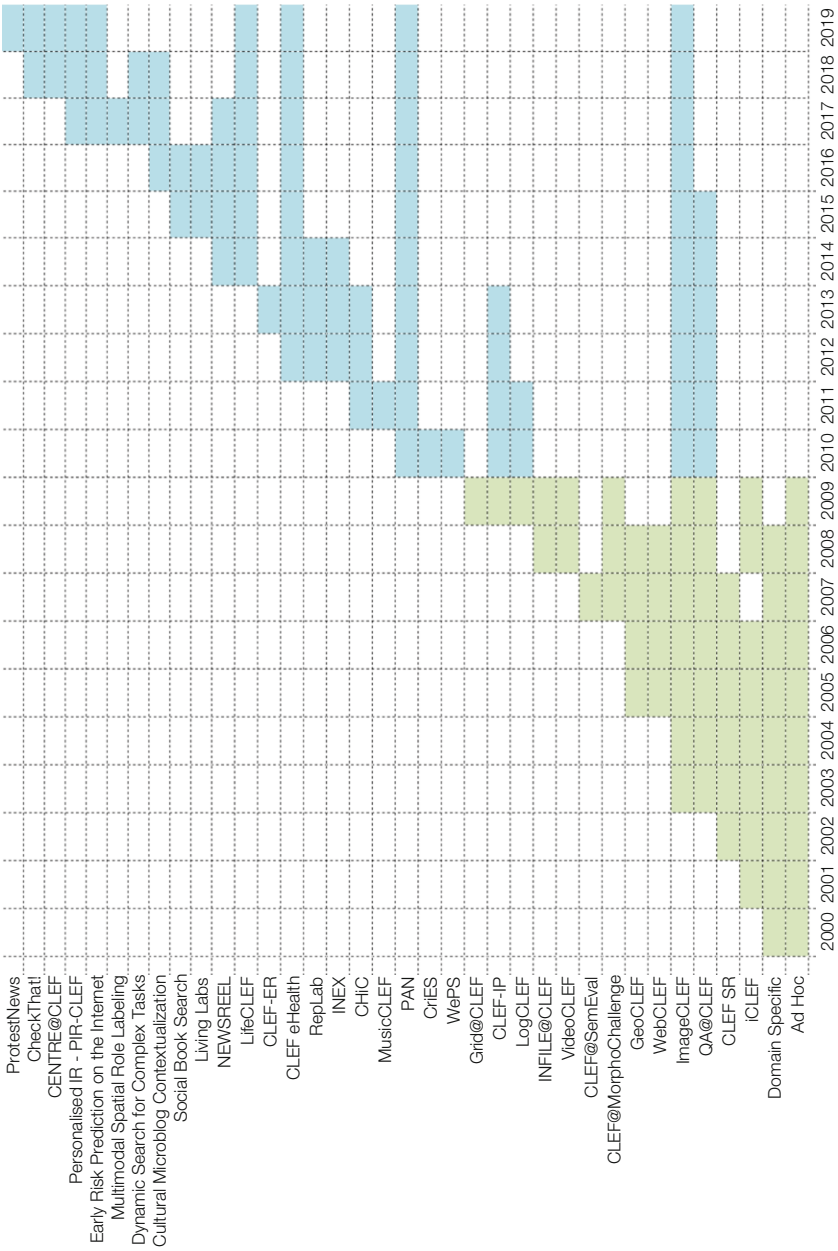


Fig. 1 Evolution of CLEF activities over time

CLEF activities (Sect. 5), and the impact that we feel that CLEF has had on research into information access and evaluation both in Europe and globally (Sect. 6).

## 1.1 Experimental Evaluation

*Information Retrieval (IR)* is concerned with developing methods, algorithms, and systems which allow users to retrieve and access digitally stored information, in whatever language and media, relevant to their needs.

In IR users express their needs by means of queries—typically keyword-based queries expressed in natural language—that are often vague and imprecise formulations of their actual information needs, and systems retrieve items—generally termed documents—that match the user query and rank them by an estimation of their relevance to the query.

Since user queries and documents can be somewhat ambiguous, since a lot of contextual and task information is often left implicit, and since the notion of relevance itself is very complex and can change as the user progresses in the search (Saracevic 1975; Mizzaro 1997), IR systems adopt a *best match* approach, where results are ranked according to how well queries can be matched against documents, but always knowing that there will be some sort of inaccuracy and fuzziness.

IR system performance can be evaluated from two different standpoints, *efficiency* and *effectiveness*. Efficiency is concerned with the algorithmic costs of IR systems, i.e. how fast they are in processing the needed information and how demanding they are in terms of the computational resources required. Effectiveness, instead, is concerned with the ability of IR systems to retrieve and properly rank relevant documents while at the same time suppressing the retrieval of non relevant ones. The ultimate goal is to satisfy the user’s information needs.

While efficiency could also be assessed formally, e.g. by proving the computational complexity of the adopted algorithms, effectiveness can be assessed only experimentally and this is why IR is a discipline strongly rooted in experimentation since its inception (Harman 2011; Spärck Jones 1981). Over the years, experimental evaluation has thus represented a main driver of progress and innovation in the IR field, providing the means to assess, understand, and improve the performance of IR systems from the viewpoint of effectiveness.

Experimental evaluation addresses a very wide spectrum of cases, ranging from system-oriented evaluation (Sanderson 2010) to user-oriented evaluation (Kelly 2009). In this volume, we will mainly focus on system-oriented evaluation which is performed according to the *Cranfield paradigm* (Cleverdon 1967).

Figure 2 summarizes the Cranfield paradigm which is based on experimental collections  $\mathcal{C} = (D, T, RJ)$  where: a corpus of documents  $D$  represents the domain of interest; a set of topics  $T$  represents the user information needs; and human-made relevance judgments  $RJ$  are the “correct” answers, or ground-truth, determining, for each topic, the relevant documents. Relevance judgments are



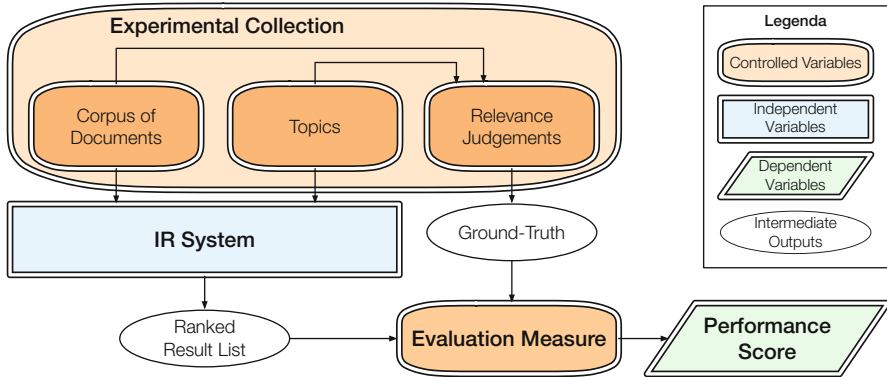


Fig. 2 The Cranfield paradigm for experimental evaluation

typically expressed as either binary relevance, i.e. relevant or not relevant, or as graded relevance (Kekäläinen and Järvelin 2002), e.g. not relevant, partially relevant, highly relevant. The ranked result lists, i.e. the IR system outputs, are then scored with respect to the ground-truth using several evaluation measures (Sakai 2014a). The evolution of Cranfield in IR system evaluation is discussed in detail in a following chapter by Ellen Voorhees.

The main goal of this experimental setup is to be able to compare the performance of different IR systems in a robust and repeatable way, as they are all scored with respect to the same experimental collection. Experimental collections and evaluation measures are *controlled variables*, since they are kept fixed during experimentation; IR systems are *independent variables*, since they are the object of experimentation, compared one against the other; and, performance scores are the *dependent variables*, since their observed value changes as IR systems change (Fuhr 2012).

Carrying out experimental evaluation according to the Cranfield paradigm is very demanding in terms of both the time and the effort required to prepare the experimental collection. Therefore, it is usually carried out in publicly open and large-scale evaluation campaigns, often at international level, as exemplified in the next section, to share the effort, compare state-of-the-art systems and algorithms on a common and reproducible ground, and maximize the impact. Tetsuya Sakai, in another chapter in this first part of the book, provides a detailed description on how to setup a Cranfield style evaluation task, create experimental collections, and use evaluation measures.

In fact, IR evaluation adopts a whole breadth of evaluation measures (Sakai 2014a) because different evaluation measures embed different user models in scanning the result list and thus represent different angles on the effectiveness of an IR system. *Average Precision (AP)* (Buckley and Voorhees 2005), *Precision at Ten (P@10)* (Büttcher et al. 2007), *Rank-Biased Precision (RBP)* (Moffat and Zobel 2008), *normalized Discounted Cumulative Gain (nDCG)* (Järvelin and

Kekäläinen 2002), and *Expected Reciprocal Rank (ERR)* (Chapelle et al. 2009) are among the most commonly adopted measures. Evaluation measures are typically studied in an empirical way, e.g. by using correlation analysis (Voorhees and Harman 1998), discriminative power (Sakai 2006, 2012), or robustness to pool downsampling (Buckley and Voorhees 2004; Yilmaz and Aslam 2006). On the other hand, few studies have been undertaken to understand the formal properties of evaluation measures and they have just scratched the surface of the problem: (Bollmann 1984; Busin and Mizzaro 2013; Amigó et al. 2013; van Rijsbergen 1974; Ferrante et al. 2015, 2017, 2019). The following chapters by Voorhees and Sakai both discuss evaluation measures in more detail.

Finally, statistical analyses and statistical significance testing play a fundamental role in experimental evaluation (Carterette 2012; Hull 1993; Sakai 2014b; Savoy 1997) since they provide us with the means to properly assess differences among compared systems and to understand when they actually matter.

The activities of CLEF, as described in the rest of this book, have been conducted within this context of theory and practice, with the results helping to stimulate progress in the field of IR system experimentation and evaluation.

## 1.2 *International Evaluation Initiatives*

There are a number of evaluation initiatives around the world that follow the Cranfield paradigm, extending and adapting it to meet local requirements. In this section, we list the major ones, indicating their relationship with CLEF.

As is described in the chapter by Voorhees, IR experimental evaluation was initiated in 1992 by the US National Institute of Standards and Technology, in the *Text REtrieval Conference (TREC)* (Harman and Voorhees 2005). The TREC conference series has constituted the blueprint for the organization of evaluation campaigns, providing guidelines and paving the way for others to follow.<sup>3</sup> In 1997, TREC included a track for *Cross-Language Information Retrieval (CLIR)*. The aim was to provide researchers with an infrastructure for evaluation that would enable them to test their systems and compare the results achieved using different cross-language strategies (Harman et al. 2001).

However, after 3 years within TREC, it was decided that Europe with its diversity of languages was better suited for the coordination of an activity that focused on multilingual aspects of IR. Not only was it far easier in Europe to find the people and groups with the necessary linguistic competence to handle the language-dependent issues involved in creating test collections in different languages, but European researchers, both in academia and industry, were particularly motivated to study the problems involved in searching over languages other than English. Consequently, with the support of TREC, the *Cross-Language Evaluation Forum (CLEF)* was

---

<sup>3</sup>See <http://trec.nist.gov/>.

launched in 2000 by a consortium with members from several different European countries, and test collections were created in four languages (English, French, German and Italian).

The decision to launch CLEF in Europe came just 1 year after the first *NII Testbeds and Community for Information access Research (NTCIR)* workshop was held in Asia.<sup>4</sup> NTCIR also saw the creation of test collections in languages other than English, i.e. in this case Asian languages, as strategic. NTCIR-1 thus included a task for cross-language Japanese to English IR and since then NTCIR has offered test collections and tasks for Chinese and Korean as well as Japanese and English. Organized on an 18 monthly cycle, NTCIR has grown steadily over the years, covering many diverse information access tasks including, but not limited to, information retrieval, question answering, text summarisation and text mining, always with an emphasis on East Asian languages. In 2017, with its twelfth conference, NTCIR celebrated its 20th birthday.

In 2006 and 2007, in response to requests from colleagues in India, CLEF organized mono- and cross-language text retrieval tasks dedicated to Indian languages. Descriptions of this activity can be found in the CLEF Workshop Proceedings for those years (Nardi et al. 2006, 2007). This preliminary action helped to lead to the birth of a new evaluation initiative in India: the *Forum for Information Retrieval Evaluation (FIRE)*<sup>5</sup> in 2008. The objective of FIRE is to stimulate the development of IR systems capable of handling the specific needs of the languages of the Indian sub-continent. When FIRE began, Indian language information retrieval research was in a relatively primitive stage (especially with regard to large-scale quantitative evaluation). FIRE has had a significant impact on the growth of this discipline by providing test collections in many Indian languages (e.g. Bengali, Gujarati, Hindi, Marathi, Tamil, Telugu) and a forum where beginners can meet with and learn from experts in the field. Over the years, FIRE has evolved to include new domains like plagiarism detection, legal information access, mixed script information retrieval and spoken document retrieval.

Another important activity, which was first launched in CLEF before becoming an independent IR evaluation initiative in 2010, is MediaEval.<sup>6</sup> MediaEval attracts participants interested in multimodal approaches to multimedia involving, e.g., speech recognition, multimedia content analysis, music and audio analysis, viewer affective response, and social networks. In particular, it focuses on the human and social aspects of multimedia tasks. MediaEval began life as VideoCLEF, a track offered in CLEF in 2008 and 2009. Relations between the two activities have been maintained and, in 2017, the MediaEval workshop and the CLEF conference were co-located and run in close collaboration. More details on MediaEval can be found in the chapter by Gareth Jones in this volume.

---

<sup>4</sup>See <http://research.nii.ac.jp/ntcir/>.

<sup>5</sup>See <http://fire.irs.res.in/>.

<sup>6</sup>See <http://www.multimediaeval.org/>.

On the other hand, the *INitiative for the Evaluation of XML Retrieval (INEX)*, run as a separate evaluation initiative from 2002 to 2011, decided in 2012 to run as a Lab under the CLEF umbrella. This Lab ran in CLEF until 2016. INEX promoted the evaluation of search engines for focused retrieval, i.e. the identification of the relevant parts of a relevant document. This can take many forms, e.g. passage retrieval from a long document, element retrieval from an XML document, page retrieval from books, as well as question answering. The chapter by Kamps et al. describes the important contribution made by INEX to experimental evaluation.

Each of the initiatives mentioned has been studied to meet the perceived needs of a specific community, reflecting linguistic, cultural and resource differences, while being designed within a common theoretical framework. This common background has facilitated discussion and exchange of ideas between the different groups and, at times, tasks run in collaboration. The aim is to avoid the duplication of effort and to provide complementary challenges, thus achieving a synergy of ideas and activities. An example of this is the CLEF/NTCIR/TREC task focused on Reproducibility, first experimented at CLEF in 2018<sup>7</sup> (Ferro et al. 2018). The objectives are to (1) reproduce the best—or most interesting—results achieved in previous editions of CLEF, NTCIR and TREC by using standard open source IR systems; and then (2) to offer the additional components and resources developed in this activity to the IR community with the aim of improving existing open source systems.

## 2 CLEF 1.0: Cross-Language Evaluation Forum (2000–2009)

When CLEF began in 2000, cross language IR had only just started to be recognized as a separate sub-discipline,<sup>8</sup> there were very few research prototypes in existence and work was almost entirely concentrated on text retrieval systems running on at most two languages. Thus, when CLEF was launched, the declared objectives were “to develop and maintain an infrastructure for the testing and evaluation of information retrieval systems operating on European languages, in both monolingual and cross-language contexts, and to create test-suites of reusable data that can be employed by system developers for benchmarking purposes” (Peters 2001). The aim was to promote the development of IR systems and tools in languages other than English and to stimulate the growth of the European research community in this area. However, while the first three editions of CLEF were dedicated to mono- and multilingual ad-hoc text retrieval, gradually the scope of activity was extended to include other kinds of text retrieval across languages (i.e., not just document

---

<sup>7</sup>See <http://www.centre-eval.org/>.

<sup>8</sup>The first workshop on “Cross-Lingual Information Retrieval” was held at the Nineteenth ACM-SIGIR Conference on Research and Development in Information Retrieval in 1996. At this meeting there was considerable discussion aimed at establishing the scope of this area of research and defining the core terminology. The first 10 years of CLEF did much to consolidate this field of study.

retrieval but question answering and geographic IR as well) and on other media (i.e., collections containing images and speech). The goal was not only to meet but also to anticipate the emerging needs of the R & D community and to encourage the development of next generation multilingual IR systems.

In this section, dedicated to CLEF 1.0, we outline the main activities undertaken in the first 10 years.

## **2.1 *Tracks and Tasks in CLEF 1.0***

Initially CLEF was very much influenced by its origins as a track within TREC. We not only adopted the same experimental paradigm that had been studied and implemented within TREC, but also inherited much of the vocabulary and the organizational framework. Therefore, for the first 10 years the different activities were run under the heading of *Tracks*. Each track was run by a coordinating group with specific expertise in the area covered.<sup>9</sup> The coordinators were responsible for the definition and organization of the evaluation activity of their *Track* throughout the year. The results were presented and discussed at the annual CLEF Workshop held in conjunction with the European Conference for Digital Libraries. Most tracks offered several different tasks and these tasks normally varied each year according to the interests of the track coordinators and participants. This meant that the number of tracks offered by CLEF 1.0 increased over the years from just two in 2000 to ten separate tracks in 2009. Activities were mostly divided into two groups: tracks concerned with text retrieval and those which studied retrieval in other media: image, speech and video. The focus was always on collections in languages other than English. In this section we present the main tracks.

Of course, some of the CLEF 1.0 tracks continued as Labs in CLEF 2.0. This is the case, for example, of ImageCLEF and CLEF-QA, two of the most popular activities, in terms both of participation and diversity of tasks. For this reason, they are presented both as tracks in this section and as Labs in Sect. 3. On the other hand, the descriptions of LogCLEF and CLEF-IP, pilot experiments at the very end of CLEF 1.0 and Labs in the following years, appear in the CLEF 2.0 section.

### **2.1.1 Multilingual Text Retrieval (2000–2009)**

Ad-Hoc document retrieval was the core track in CLEF 1.0. It was the one track that was offered every year and was considered of strategic importance. For this reason, we describe it in some detail here. The aim of the track was to promote

---

<sup>9</sup>It is impossible to acknowledge all the researchers and institutions that have been involved in the coordination of CLEF. Many, but certainly not all, are represented by the authors of the papers in this volume.

the development of monolingual and cross-language text retrieval systems. From 2000–2007, the track exclusively used target collections of European newspaper and news agency documents and worked hard at offering increasingly complex and diverse tasks, adding new languages every year. Up until 2005, European languages were also used for the queries. In 2006 and 2007, in a collaboration with colleagues from the *Information Retrieval Society of India (IRSI)* which would lead in 2008 to the launching of FIRE, we added the possibility to query the English document collection with queries in a number of Indian languages. In 2008 and 2009, as a result of a joint activity with the Database Research Group of Tehran University, we included a test collection in Farsi, the Hamshahri corpus of 1996–2002 newspapers. Monolingual and cross-language (English to Persian) tasks were offered. As was to be expected, many of the eight participants focused their attention on problems of stemming. Only three submitted cross-language runs.

The addition of queries and a document collection in non European languages was important as it provided the opportunity to test retrieval systems on languages with very different scripts and syntactic structures. For example, the decision to offer a Persian target collection was motivated by several reasons: the challenging script (a modified version of Arabic with elision of short vowels) written from right to left; the complex morphology (extensive use of suffixes and compounding); the political and cultural importance.

In 2006 we added a task designed for more experienced participants, the so-called “robust task”, which used test collections from previous years in six languages (Dutch, English, French, German, Italian and Spanish) with the objective of rewarding experiments which achieve good stable performance over all queries rather than high average performance.

In 2008 we also introduced a task offering monolingual and cross-language search on library catalog records. It was organized in collaboration with The European Library (TEL)<sup>10</sup> and used three collections from the catalogs of the British Library, the Bibliothèque Nationale de France and the Austrian National Library. The underlying aim was to identify the most effective retrieval technologies for searching this type of very sparse multilingual data. In fact, the collections contained records in many languages in addition to English, French or German. The task presumed a user with a working knowledge of these three languages who wants to find documents that can be useful for them in one of the three target catalogs. Records in other languages were counted irrelevant. This was a challenging task but proved popular; participants tried various strategies to handle the multilinguality of the catalogs. The fact that the best results were not always obtained by experienced CLEF participants shows that the traditional approaches used for newspaper document retrieval are not necessarily the most effective for this type of data. The task was offered for 2 years.

Another task, offered for just 2 years, was designed to attract participation from groups interested in *Natural Language Processing (NLP)*. English test data from

---

<sup>10</sup>See <http://www.theeuropeanlibrary.org/>.

previous years was used but the organizers provided *Word Sense Disambiguation (WSD)* for documents and queries. Both monolingual and bilingual (Spanish to English) tasks were activated. This task ran for 2 years, however, the results were inconclusive. Overall, little or no improvement in performance was achieved by groups attempting to exploit the WSD information.

The focus of the Ad-Hoc track on multilingual IR implied considering and understanding the challenges posed to information access technology by variation between languages in their writing systems, and in their morphological, syntactic and lexical properties. This problematic is investigated in the chapter by Karlgren et al. in Part III of this volume.

Table 1 gives a detailed breakdown of the collections and tasks offered for Ad-Hoc in each of these 10 years. It can be seen that bilingual tasks were often proposed for unusual pairs of languages, such as Finnish to German, or French to Dutch. In addition multilingual tasks were offered in which queries in one language were posed to target collections in a varying number of languages.  $x$  as the query language in the bilingual and multilingual tasks denotes any of the languages offered for the monolingual task of that year.

The results of this track were considerable. It is probably true to say that it has done much to foster the creation of a strong European research community in the CLIR area. It provided the resources, the test collections and also the forum for discussion and comparison of ideas and results. Groups submitting experiments over several years showed flexibility in advancing to more complex tasks, from monolingual to bilingual and multilingual experiments. Much work was done on fine tuning for individual languages while other efforts concentrated on developing language independent strategies (McNamee and Mayfield 2004). Over the years, there was substantial proof of significant increase in retrieval effectiveness in multilingual settings by systems of CLEF participants (Braschler 2004).

The paper by Savoy and Braschler in this volume discusses some of the lessons learnt from this track.

### 2.1.2 The Domain-Specific Track (2001–2008)

Another text retrieval track offered for many years in CLEF 1.0 was the Domain-Specific track which was organised by a group with specific expertise in the area covered.<sup>11</sup> Mono- and cross-language retrieval was investigated using structured data (e.g. bibliographic data, keywords and abstracts) from scientific reference databases. The track used German, English and Russian target collections in the social science domain. A multilingual controlled vocabulary was also provided. A main finding was that metadata-based search can achieve similar results as those

---

<sup>11</sup>This track was coordinated by Michael Kluck, Informationszentrum Sozialwissenschaften (IZ), Germany.

**Table 1** CLEF 2000–2009 ad-hoc tasks

| Edition   | Monolingual  | Bilingual  | Multilingual  |
|-----------|--|--|---|
| CLEF 2000 | de; fr; it   | x → en   | x → de; en; fr; it  |
| CLEF 2001 | de; es; fr; it; nl                                 | x → en<br>x → nl   | x → de; en; es; fr; it  |
| CLEF 2002 | de; es; fi; fr; it; nl; sv                         | x → de; es; fi; fr; it; nl; sv<br>x → en (newcomers only)                                    | x → de; en; es; fr; it  |
| CLEF 2003 | de; es; fi; fr; it; nl; ru; sv                     | it → es<br>de → it<br>fr → nl<br>fi → de<br>x → ru<br>x → en (newcomers only)                | x → de; en; es; fr<br>x → de; en; es; fi; fr; it; nl; sv          |
| CLEF 2004 | fi; fr; ru; pt                                     | es; fr; it ;ru → fi<br>de; fi; nl; sv → fr<br>x → ru<br>x → en (newcomers only)              | x → fi; fr; ru; pt  |
| CLEF 2005 | bg; fr; hu; pt                                     | x → bg; fr; hu; pt   | Multi8 2yrson (as in CLEF 2003)<br>Multi8 Merge (as in CLEF 2003) |
| CLEF 2006 | bg; fr; hu; pt<br>Robust<br>de; en; es; fr; it; nl | x → bg; fr; hu; pt<br>am; hi; id; te; or → en<br>Robust<br>it → es<br>fr → nl<br>en → de     | Robust<br>x → de; en; es; fr; it; nl                              |
| CLEF 2007 | bg; cz; hu<br>Robust<br>en; fr; pt                 | x → bg; cz; hu<br>am; id; or; zh → en<br>bn; hi; mr; ta; te → en<br>Robust<br>x → en; fr; pt |   |
| CLEF 2008 | fa<br>TEL<br>de; en; fr<br>Robust WSD<br>en        | en → fa<br>TEL<br>x → de; en; fr<br>Robust WSD<br>es → en                                    |   |
| CLEF 2009 | fa<br>TEL<br>de; en; fr<br>Robust WSD<br>en        | en → fa<br>TEL<br>x → de; en; fr<br>Robust WSD<br>es → en                                    |   |

The following ISO 639-1 language codes have been used: *am* Amharic, *bg* Bulgarian, *bn* Bengali, *de* German, *en* English, *es* Spanish, *fa* Farsi, *fi* Finnish, *fr* French, *hi* Hindi, *hu* Hungarian, *id* Indonesian, *it* Italian, *mr* Marathi, *nl* Dutch, *or* Oromo, *pt* Portuguese, *ru* Russian, *sv* Swedish, *ta* Tamil, *te* Telugu. *TEL* data from The European Library



obtained using full-text. The results of the mono- and cross-language experiments were very similar in terms of performance to those achieved in the ad-hoc track.

In CLEF 2.0, domain-specific activities acquired a multimedia/multimodal perspective and included tasks involving patent retrieval, health management and biodiversity.

### 2.1.3 Interactive Cross-Language Retrieval (2002–2009)

In the iCLEF track, cross-language search capabilities were studied from a user-inclusive perspective. A central research question was how best to assist users when searching information written in unknown languages, rather than how best an algorithm can find information written in languages different from the query language. In 2006, iCLEF moved from the news collections used in the ad-hoc tasks in order to explore user behaviour in a collection where the cross-language search necessity arises more naturally for average users. The choice fell on Flickr, a large-scale, online image database based on an extensive social network of WWW users, with the potential for offering both challenging and realistic multilingual search tasks for interactive experiments. The search interface provided by the iCLEF organizers was a basic cross-language retrieval system for the Flickr image database<sup>12</sup> presented as an online game: the user was given an image, and had to find it again without any a priori knowledge of the language(s) in which the image is annotated. The game was publicized on the CLEF mailing list and prizes were offered for the best results in order to encourage participation. The main novelty of the iCLEF 2008 experiments was the shared analysis of a search log from a single search interface provided by the organizers (i.e. the focus was on log analysis, rather than on system design).

The 2008 experiments resulted in a truly reusable data set (the first time in iCLEF!), with 5000 complete search sessions recorded and 5000 post-search and post-experience questionnaires. 200 users from 40 countries played an active role in these experiments which covered six target languages. A main observation was that, in addition to better CLIR algorithms, more research was needed on interactive features to help users bridge the language gap.

The track was organised in a similar way in 2009. The organizers provided a default multilingual search system which accessed images from Flickr, with the whole iCLEF experiment run as an online game. Interaction by users with the system was recorded in log files which were shared with participants for further analyses, and provide a future resource for studying various effects on user-orientated cross-language search.

---

<sup>12</sup>See <http://www.flickr.com/>.

### **2.1.4 The Question-Answering Track (2003–2015)**

From 2003 on, CLEF also offered mono- and cross-language question answering tasks. The QA track was instrumental in encouraging researchers working in the natural language processing field to participate in CLEF. The main scenario in the early years was event targeted QA on a heterogeneous document collection. Besides the usual news collections used in the ad-hoc track, articles from Wikipedia were also considered as sources of answers and parallel aligned European legislative documents were included from 2009.

This track was inspired by the work in TREC on question answering but in CLEF the focus was on multilinguality. Many monolingual and cross-language sub-tasks were offered: Basque, Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish were proposed as both query and target languages; not all were used in the end. This track proved very popular in CLEF 1.0 and was, in fact, continued in CLEF 2.0. Over the years, a lot of resources and know-how were accumulated. One important lesson learnt was that offering so many language possibilities meant that there were always only a few systems participating in the same task, with the same languages. This meant that comparative analysis was often problematic. The chapter by Peñas et al. in this volume discusses in detail the design, experience and results of question answering activities in CLEF.

### **2.1.5 Cross-Language Retrieval in Image Collections (2003–2019)**

Although at the beginning CLEF was very much focused on text retrieval, in 2003 it was decided to offer a track testing the retrieval of images from multilingual collections. ImageCLEF was thus launched with the goal of providing support for the evaluation of (1) multilingual image retrieval methods, to compare the effect of retrieval of image annotations and query formulations in several languages, (2) multimodal information retrieval methods based on the combination of visual and textual features, and (3) language-independent methods for the automatic annotation of images with concepts. The initial activity in this track is described in the chapter by Clough and Tsirikia in this volume. However, over the years, the track became increasingly complex. With the introduction of search on medical images in CLEF 2004, it also became very oriented towards the needs of an important user community (see the chapter by Müller et al.).

ImageCLEF rapidly became the most popular track in CLEF 1.0, even though (or maybe because) it was the track that deals the least with language and linguistic issues. This interest was to continue and diversify in CLEF 2.0. This is also exemplified in the chapters by Wang et al. and Piras et al.

### **2.1.6 Spoken Document/Speech Retrieval (2003–2007)**

Following a preliminary investigation carried out as part of the CLEF 2002 campaign, a Cross-Language Spoken Document Retrieval (CLSDR) track was organized in CLEF 2003 and 2004. The track took as its starting point automatic transcripts prepared by NIST for the TREC 8-9 SDR tracks and generated using different speech recognition systems. The task consisted of retrieving news stories within a repository of about 550h of transcripts of American English news. The original English short search topics were formulated in French and German, to provide a CL-SDR task.

The CLEF 2005 Cross-Language Speech Retrieval (CL-SR) track followed these 2 years of experimentation but used audio data from the MALACH (Multilingual Access to Large Spoken Archives) collection which is based on interviews with Holocaust survivors from the archives of the Shoah Visual History Foundation. Spontaneous, conversational speech lacks clear topic boundaries and is considerably more challenging for the Automatic Speech Recognition, (or ASR), techniques on which fully-automatic content-based search systems are based. Although, advances in ASR had made it possible to contemplate the design of systems that would provide a useful degree of support for searching large collections of spontaneous conversational speech, no representative test collection that could be used to support the development of such systems was widely available for research use at that time. The principal goal of the CLEF CL-SR track was thus to create such a test collection. The data used was mainly in English and Czech. Topics were developed in several languages. Additional goals included benchmarking the current state of the art for ranked retrieval of spontaneous conversational speech and fostering interaction among a community of researchers with interest in that challenge.

Those goals were achieved. Over 3 years, research teams from 14 universities in 6 countries submitted runs for official scoring. The resulting English and Czech collections are the first information retrieval test collections of substantial size for spontaneous conversational speech. Unique characteristics of the English collection fostered research comparing searches based on automatic speech recognition and manually assigned metadata, and unique characteristics of the Czech collection inspired research on evaluation of information retrieval from unsegmented speech.

The CLEF spoken document and speech retrieval activities are described in more detail in the chapter by Gareth Jones.

### **2.1.7 Multilingual Web Retrieval (2005–2008)**

The WebCLEF track focused on evaluation of systems providing multi- and cross-lingual access to web data. In the final year, a multilingual information synthesis task was offered, where, for a given topic, participating systems were asked to extract important snippets from web pages (fetched from the live web and provided by the task organizers). The systems had to focus on extracting, summarizing, filtering and presenting information relevant to the topic, rather than on large scale

web search and retrieval per se. The aim was to refine the assessment procedure and evaluation measures. WebCLEF 2008 had lots of similarities with (topic-oriented) multidocument summarization and with answering complex questions. An important difference was that at WebCLEF, topics could come with extensive descriptions and with many thousands of documents from which important facts had to be mined. In addition, WebCLEF worked with web documents, which can be very noisy and redundant.

Although the Internet would seem to be the obvious application scenario for a CLIR system, WebCLEF had a rather disappointing participation. For this reason, the track was dropped.

### **2.1.8 Geographical Retrieval (2005–2008)**

The purpose of GeoCLEF was to test and evaluate cross-language geographic information retrieval for topics with a geographic specification. How best to transform into a machine readable format the imprecise description of a geographic area found in many user queries was considered an open research problem. This track was run for 4 years in CLEF, examining geographic search of a text corpus. Some topics simulated the situation of a user who poses a query when looking at a map on the screen. In GeoCLEF 2006 and 2007, it was found that keyword based systems often do well on the task and the best systems worked without any specific geographic resource. In 2008 the best monolingual systems used specific geo reasoning; there was much named-entity recognition (often using Wikipedia) and NER topic parsing. Geographic ontologies were also used (such as GeoNames and World Gazetteer), in particular for query expansion.

The track was coordinated by Frederic Gey and Ray Larson of UC Berkeley, School of Information. In 2009, they decided to move this activity from Europe to Asia and initiated a geotemporal retrieval task at NTCIR-8. However, in CLEF 2009, a new track, LogCLEF, continued to study information retrieval problems from the geographical perspective (see Sect. 3.1.9).

### **2.1.9 Multilingual Information Filtering (2008–2009)**

The purpose of the INFILE (INformation FILtering & Evaluation) track, sponsored by the French National Research Agency, was to evaluate cross-language adaptive filtering systems. The goal of these systems is to successfully separate relevant and non-relevant documents in an incoming stream of textual information with respect to a given profile. The document and profile may be written in different languages.

INFILE extended the last filtering track of TREC 2002 in the following ways:

- Monolingual and cross-language tasks were offered using a corpus of 100,000 Agence France Press (AFP) comparable newswire stories for Arabic, English and French;

- Evaluation was performed by an automatic querying of test systems with a simulated user feedback. A curve of the evolution of efficiency was computed along with more classical measures already tested in TREC.

Unfortunately, the innovative crosslingual aspect of the task was not really explored, since most of the runs were monolingual English and no participant used the Arabic topics or documents.

### 2.1.10 Cross-Language Video Retrieval (2008–2009)

The aim of the VideoCLEF track was to develop and evaluate tasks related to analysis of and access to multilingual multimedia content. Participants used a video corpus containing episodes of a dual language television program in Dutch and English, accompanied by speech recognition transcripts. The dual language programming of Dutch TV offered a unique scientific opportunity, presenting the challenge of how to exploit speech features from both languages.

In 2010, the VideoCLEF organisers decided to set up an independent benchmarking initiative, known as MediaEval.<sup>13</sup> MediaEval attracts participants who are interested in multimodal approaches to multimedia involving, e.g., speech recognition, multimedia content analysis, music and audio analysis, user-contributed information (tags, tweets), viewer affective response, social networks, temporal and geo-coordinates. This initiative is having a lot of success with a very active participation. Results are presented in an annual workshop.

More information on VideoCLEF and MediaEval is given in the chapter by Gareth Jones.

### 2.1.11 Component-Based Evaluation (2009)

Grid@CLEF was a pilot experiment focused on *component-based evaluation* and aimed at establishing a long term activity comprising a series of systematic experiments in order to improve the comprehension of *MultiLingual Information Access (MLIA)* systems and gain an exhaustive picture of their behaviour with respect to languages. To this end, Grid@CLEF introduced the notion of *Grid of Points (GoP)* (Ferro and Harman 2010), i.e. a set of IR systems originated by all the possible combinations of components under experimentation.

Grid@CLEF 2009 offered traditional monolingual ad-hoc tasks in 5 different languages (Dutch, English, French, German, and Italian) and used consolidated and very well known collections from CLEF 2001 and 2002 with a set of 84 topics. Participants had to conduct experiments according to the *Coordinated Information Retrieval Components Orchestration (CIRCO)* framework, an XML-based protocol

---

<sup>13</sup>See <http://www.multimediaeval.org/>.

which allows for a distributed, loosely coupled, and asynchronous experimental evaluation of IR systems. A Java library was provided which could be exploited to implement CIRCO together with an example implementation with the Lucene IR system. The task proved to be particularly challenging. Of the 9 original participants, only 2 were able to submit runs. They used different IR systems or combination of them, namely Lucene, Terrier, and Cheshire II. Partly because it was seen as overly complex, the activity was suspended.

Even if only run for 1 year, Grid@CLEF seeded some follow-up research lines. The interest in component-based evaluation was continued by Hanbury and Müller (2010) and embedded in the idea of evaluation-as-a-service (Hopfgartner et al. 2018), as discussed in a chapter in this book by Hanbury and Müller. The idea of GoP was taken up by Ferro and Silvello (2016, 2017) to develop *ANalysis Of Variance (ANOVA)* models able to break-down overall system performance into those of the constituting components. GoP have also been exploited by Angelini et al. (2018) to develop a *Visual Analytics (VA)* system to explore and intuitively make sense of them, as is described in a chapter in this book by Ferro and Santucci.

### 3 CLEF 2.0: Conference and Labs of the Evaluation Forum (2010–2019)

The second period of CLEF started with a clear and compelling question: after a successful decade studying multilinguality for European languages, what were the main unresolved issues currently facing us? To answer this question, we turned to the CLEF community to identify the most pressing challenges and to list the steps to be taken to meet them.

The discussion led to the definition and establishment of the *CLEF Initiative*, whose main mission is to promote research, innovation, and the development of information access systems with an emphasis on multilingual and *multimodal* information with various levels of structure.

In the CLEF Initiative an increased focus is on the *multimodal* aspect, intended not only as the ability to deal with information coming in multiple media but also in different modalities, e.g. the Web, social media, news streams, specific domains and so on. These different modalities should, ideally, be addressed in an integrated way; rather than building vertical search systems for each domain/modality the interaction between the different modalities, languages, and user tasks needs to be exploited to provide comprehensive and aggregated search systems. Thus, multimodality became a major theme of CLEF 2.0.

The new challenges for CLEF also called for a renewal of its structure and organization. The annual CLEF meeting is no longer a Workshop, held in conjunction with the European Digital Library Conference (ECDL, now TPDL), but has become an independent event, held over 3.5–4 days and made up of two interrelated activities: the *Conference* and the workshops of the *Labs*.

The *Conference* is a peer-reviewed conference, open to the IR community as a whole and not just to *Lab* participants, and aims at stimulating discussion on innovative evaluation methodologies and fostering a deeper analysis and understanding of experimental results. The *Labs* replace the *Tracks* of CLEF 1.0 and are organised on a yearly basis, culminating with the annual meeting where the results are discussed. Lab coordinators are responsible for the organization of the IR system evaluation activities of their Lab throughout the year and for their annual Lab workshop. They also give plenary Lab “overview presentations” during the conference to allow non-participants to get a sense of the direction of the research frontiers. The *Conference* and the *Labs* are expected to interact, bringing new interests and new expertise into CLEF.

Moreover, in order to favour participation and the introduction of new perspectives, CLEF now has an open-bid process which allows research groups and institutions to bid to host the annual CLEF event and to propose new themes, characterizing each edition.

The new challenges and new organizational structure motivated the change in name for CLEF: from the *Cross-Language Evaluation Forum* to *Conference and Labs of the Evaluation Forum*, in order to reflect the widened scope.

### **3.1 Workshops and Labs in CLEF 2.0**

The move from the Tracks of CLEF 1.0 to the Labs of CLEF 2.0 was first made in CLEF 2010. A procedure was set up for the selection of the Labs to be held each year. A Lab Selection Committee launches a Call for Proposals in the Fall of the previous year. Proposals are accepted for two different types of Labs:

- Benchmarking Labs, providing a “campaign-style” evaluation for specific information access problems, similar in nature to the traditional CLEF campaign “Tracks” of CLEF 1.0. Topics covered by campaign-style labs can be inspired by any information access-related domain or task.
- Workshop-style Labs, following a more classical “workshop” pattern, exploring issues of evaluation methodology, metrics, processes etc. in information access and closely related fields, such as natural language processing, machine translation, and human-computer interaction.

For first time proposers, it is highly recommended that a lab workshop be first organised to discuss the format, the problem space, and the practicalities of the shared task. At the annual meeting, Labs are organised so that they contain ample time for general discussion and engagement by all participants—not just those presenting campaign results and papers. The criteria adopted for selection of Lab proposals include: importance of problem, innovation, soundness of methodology, clear movement along a growth path, likelihood that the outcome would constitute a significant contribution to the field. Additional factors are minimal overlap with

other evaluation initiatives and events, vision for a potential continuation, and possible interdisciplinary character.

In this section, we provide a brief description of the Workshops and the Labs held in the second decade of CLEF, and shown in Fig. 1 at page 5. For completeness, we have also included indication of the activities underway in 2019. We begin by describing the 1-year experimental Workshops and continue with presentations of the fully-fledged Labs.

### **3.1.1 Web People Search (2010)**

The WePS workshop focused on person name ambiguity and person attribute extraction from Web pages and on online reputation management for organizations. The first edition of this workshop, WePS-1, was run as a Semeval 1 task in 2007, whereas WePS-2 was a workshop at the WWW 2009 Conference. WePS-1 addressed only the name co-reference problem, defining the task as clustering of web search results for a given person name. In WePS-2 the evaluation metrics were refined and an attribute extraction task for web documents returned by the search engine for a given person name was added.

In the edition of WePS at CLEF both problems were merged into a single task, where the system must return both the documents and the attributes for each of a number of people sharing a given name. This was not a trivial step from the point of view of evaluation: a system may correctly extract attribute profiles from different URLs but then incorrectly merge these profiles. While WePS-1 and WePS-2 had focused on consolidating a research community around the problem and developing an appropriate evaluation methodology, in WePS-3 the focus was on involving industrial stakeholders in the evaluation campaign, as providers of input to the task design phase and also as providers of realistic scale datasets. Intelius, Inc.—one of the main Web People Search services, providing advanced people attribute extraction and profile matching from web pages—collaborated in the activity. The discussions at this workshop resulted in the setting up of RepLab, described in Sect. 3.1.14.

### **3.1.2 Cross-Lingual Expert Search (2010)**

CriES was run as a brainstorming workshop and addressed the problem of multilingual expert search in social media environments. The main topics were multilingual expert retrieval methods, social media analysis with respect to expert search, selection of datasets and evaluation of expert search results. Online communities generate major economic value and form pivotal parts of corporate expertise management, marketing, product support, product innovation and advertising. In many cases, large-scale online communities are multilingual by nature (e.g. developer networks, corporate knowledge bases, blogospheres, Web 2.0 portals). Nowadays, novel solutions are required to deal with both the complexity of large-scale social



networks and the complexity of multilingual user behavior. It thus becomes more important to efficiently identify and connect the right experts for a given task across locations, organizational units and languages. The key objective of the workshop was to consider the problem of multilingual retrieval in the novel setting of modern social media leveraging the expertise of individual users.

### 3.1.3 Music Information Retrieval (2011)

MusiCLEF was run as a brainstorming workshop promoting the development of new methodologies for music access and retrieval on real public music collections. A major focus was on multimodal retrieval achieved by combining content-based information, automatically extracted from music files, with contextual information, provided by users via tags, comments, or reviews. MusiCLEF aimed at maintaining a tight connection with real world application scenarios, focusing on issues related to music access and retrieval that are faced by professional users. Two benchmarking tasks were studied: the automatic categorization of music to be used as soundtrack for TV shows; the automatic identification of the pieces in a music digital library. In 2012, this activity continued as part of the MediaEval Initiative,<sup>14</sup> described in Sect. 2.1.10.

### 3.1.4 Entity Recognition (2013)

The identification and normalisation of biomedical entities in scientific literature has a long tradition and a number of challenges have contributed to the development of reliable solutions. Increasingly, patient records are processed to align their content with other biomedical data resources, but this approach requires analysing documents in different languages across Europe.

CLEF-ER was a brainstorming workshop on the multilingual annotation of named entities and terminology resource acquisition with a focus on entity recognition in biomedical text in different languages and on a large scale. Several corpora in different languages, i.e. Medline titles, European Medicines Agency documents and patent claims, were provided to enable ER in parallel documents. Participants were asked to annotate entity mentions with concept unique identifiers (CUIs) in the documents of their preferred non-English language. The evaluation determined the number of correctly identified mentions against a silver standard and performance measures for the identification of CUIs in the non-English corpora. Participants could make use of the prepared terminological resources for entity normalisation and the English silver standard corpora (SSCs) as input for concept candidates in the non-English documents. Participants used different approaches including translation

---

<sup>14</sup>See <http://www.multimediaeval.org/>.

techniques and word or phrase alignments as well as lexical look-up and other text mining techniques.

### 3.1.5 Multimodal Spatial Role Labeling (2017)

The extraction of spatial semantics is important in many real-world applications such as geographical information systems, robotics and navigation, semantic search, etc. This workshop studied how spatial information could be best extracted from free text while exploiting accompanying images. The task investigated was a multimodal extension of a spatial role labeling task previously introduced in the SemEval series. The multimodal aspect of the task made it appropriate for CLEF 2.0.

### 3.1.6 Extracting Protests from News (2019)

ProtestNews aimed at testing and improving state-of-the-art generalizable machine learning and natural language processing methods for text classification and information extraction on English news from multiple countries such as India and China in order to create comparative databases of contentious political events (riots, social movements), i.e. the repertoire of contention that can enable large scale comparative social and political science studies. Three tasks were investigated: *Task 1—News article classification as protest vs. non-protest*: given a random news article, to what extent can we predict whether it is reporting a contentious politics event that has happened or is happening? *Task 2—Event sentence detection*: given a news article that is classified as positive in Task 1, to what extent can we identify the sentence(s) that contain the event information? *Task 3—Event extraction*: given the event sentence that is identified in Task 2, to what extent can we extract key event information such as place, time, participants, etc.?

### 3.1.7 Question Answering (2003–2015)

As described in the previous section, question answering was an important activity in CLEF from 2003. The QA@CLEF track, which became a Lab in 2010, examined several aspects of question answering in a multilingual setting on document collections ranging from news, legal documents, medical documents, and linked data. From 2010 on, it was decided that if progress was to be made a substantial change was needed in the design of the QA system architecture, with particular regard to answer validation and selection technologies. For this reason, the new formulation of the task after 2010 left the retrieval step aside to focus on the development of technologies able to work with a single document, answering questions about it and using the reference collections as sources of background knowledge that help the answering process. See the chapter by Peñas et al. in this volume for a more exhaustive description.

### 3.1.8 Image Retrieval (2003–2019)

As has already been stated, since its beginnings, ImageCLEF has been one of the most popular activities at CLEF. It has had the important merit of helping to make CLEF truly multidisciplinary by bringing the image processing community into close contact with researchers working on all kinds of text retrieval and in natural language processing. The main goal of the ImageCLEF Labs in CLEF 2.0 is to support multilingual users from a global community accessing an ever growing body of visual information. The objective is to promote the advancement of the fields of visual media analysis, indexing, classification, and retrieval, by developing the necessary infrastructure for the evaluation of visual information retrieval systems operating in monolingual, cross-language and language-independent contexts. ImageCLEF aims at providing reusable resources for such benchmarking purposes.

The chapters by Wang et al., Piras et al., and Müller et al. in this volume give an account of the wide range of ImageCLEF activities in CLEF 2.0.

### 3.1.9 Log File Analysis (2009–2011)

Search logs are a means to study user information needs and preferences. Interactions between users and information access systems can be analyzed and studied to gather user preferences and to learn what the user likes the most, and to use this information to personalize the presentation of results. The literature of log analysis of information systems shows a wide variety of approaches to learning user preferences by looking at implicit or explicit interaction. However, there has always been a lack of availability and use of log data for research experiments which makes the verifiability and repeatability of experiments very limited. LogCLEF investigated the analysis and classification of queries in order to understand search behavior in multilingual contexts and ultimately to improve search systems by offering openly-accessible query logs from search engines and digital libraries. An important long-term aim of the LogCLEF activity was to stimulate research on user behavior in multilingual environments and promote standard evaluation collections of log data.

Between 2009 and 2011, LogCLEF released collections of log data with the aim of verifiability and repeatability of experiments. During the three editions of LogCLEF, different collections of log datasets were distributed to the participants together with manually annotated query records to be used as a training or test set. In the final edition, a Web based interface to annotate log data was designed and created on the basis of the experience of past participants for different tasks: language identification, query classification, and query drift. The public distribution of the datasets and results and the exchange of system components aimed at advancing the state of the art in this research area (Di Nunzio et al. [2011](#)).

### 3.1.10 Intellectual Property in the Patent Domain (2009–2013)

The patent system is designed to encourage disclosure of new technologies and novel ideas by granting exclusive rights on the use of inventions to their inventors, for a limited period of time. An important requirement for a patent to be granted is that the invention it describes is novel. That is, there is no earlier patent, publication or public communication of a similar idea. To ensure the novelty of an invention, patent offices as well as other Intellectual Property (IP) service providers perform thorough searches called ‘prior art searches’ or ‘validity searches’. Since the number of patents in a company’s patent portfolio affects the company market value, well-performed prior art searches that lead to solid, difficult to challenge patents are of high importance.

The CLEF-IP Lab, which began as an experimental track at the end of CLEF 1.0, focused on various aspects of patent search and intellectual property search in a multilingual context using the MAREC collection of patents, gathered from the European Patent Office. In its first year, CLEF-IP organized one task only, a text oriented retrieval that modeled the “Search for Prior Art” done by experts at patent offices. In terms of retrieval effectiveness the results of this initial study were hard to evaluate: it appeared that the effective combination of a wide range of indexing methods produced the best results. It was agreed that further studies were needed to understand what methodology maps best to what makes a good (or better) system from the point of view of patent searchers. In the following years, the types of CLEF-IP tasks broadened to include patent text classification, patent image retrieval and classification, and (formal) structure recognition. With each task, the test collection was extended to accommodate the additional tasks.

The activity of this Lab and the results achieved are described in the chapter by Piroi and Hanbury in this volume.

### 3.1.11 Digital Text Forensics (2010–2019)

Since its first introduction in 2010, the PAN Lab has been extremely popular with a large participation. Over the years, the Lab has offered a range of tasks focusing on the general area of “Uncovering Plagiarism, Authorship and Social Software Misuse” in a multilingual context. In 2016, the Lab changed its name to the more general “Digital Text Forensics”. PAN is also a good example of the cooperation between the different international evaluation initiatives listed in Sect. 1.2. The Lab coordinators have collaborated for a number of years in the organization of evaluation tasks at *Forum for Information Retrieval Evaluation (FIRE)*, organized by the Information Retrieval Society of India, in Indian languages, Arabic and Persian.

Details on the diverse activities of this Lab are presented in the chapter by Rosso et al. in this volume.

### 3.1.12 Cultural Heritage in CLEF (2011–2013)

Cultural heritage collections preserved by archives, libraries, museums and other institutions are often multilingual and multimedia (e.g. text, photographs, images, audio recordings, and videos), usually described with metadata in multiple formats and of different levels of complexity. Cultural heritage institutions have different approaches to managing information and serve diverse user communities, often with specialized needs. The targeted audience of the CHiC lab and its tasks were developers of cultural heritage information systems, information retrieval researchers specializing in domain-specific (cultural heritage) and/or structured information retrieval on sparse text (metadata), and semantic web researchers specializing in semantic enrichment with LOD data. Evaluation approaches (particularly system-oriented evaluation) in this domain have been fragmentary and often non-standardized.

CHiC began with a brainstorming workshop in 2011 aimed at moving towards a systematic and large-scale evaluation of cultural heritage digital libraries and information access systems. In a pilot lab in 2012, a standard ad-hoc information retrieval scenario was tested together with two use-case-based scenarios (diversity task and semantic enrichment task). The 2013 lab diversified and became more realistic in its task organization. The pilot lab in 2012 demonstrated that in cultural heritage information systems ad-hoc searching might not be the prevalent form of access to this type of content. The 2013 CHiC lab focused on multilinguality in the retrieval tasks (up to 13 languages) and added an interactive task, where different usage scenarios were tested. CHiC teamed up with Europeana,<sup>15</sup> Europe’s largest digital library, museum and archive for cultural heritage objects to provide a realistic environment for experiments. Europeana provided the document collection (digital representations of cultural heritage objects) and queries from their query logs.

### 3.1.13 Retrieval on Structured Datasets (2012–2014)

Traditional IR focuses on pure text retrieval over “bags of words” but the use of structure—such as document structure, semantic metadata, entities, or genre/topical structure is of increasing importance on the Web and in professional search. INEX was founded as the INitiative for the Evaluation of XML Retrieval and has been pioneering the use of structure for focused retrieval since 2002. It joined forces with CLEF in 2012 and continued this activity. From 2015 it merged into the Social Book Search Lab (see Sect. 3.1.19). A chapter by Kamps et al. in this volume discusses INEX activities.

---

<sup>15</sup><http://www.europeana.eu>.

### 3.1.14 Online Reputation Management (2012–2014)

Reputation management is an essential part of corporate communication. It comprises activities aiming at building, protecting and repairing the images of people, organizations, products, or services. It is vital for companies (and public figures) to maintain their good name and preserve their “reputation capital”. Current technology applications provide users with a wide access to information, enabling them to share it instantly and 24 h a day due to constant connectivity. Information, including users’ opinions about people, companies or products, is quickly spread over large communities. In this setting, every move of a company, every act of a public figure are subject, at all times, to the scrutiny of a powerful global audience. The control of information about public figures and organizations at least partly has moved from them to the users and consumers. For effective Online Reputation Management (ORM) this constant flow of online opinions needs to be watched. While traditional reputation analysis is mostly manual, online media allow to process, understand and aggregate large streams of facts and opinions about a company or individual. In this context, Natural Language Processing and text mining software play key, enabling roles. Although opinion mining has made significant advances in the last few years, most of the work has been focused on products. However, mining and interpreting opinions about companies and individuals is, in general, a much harder and less understood problem, since unlike products or services, opinions about people and organizations cannot be structured around any fixed set of features or aspects, requiring a more complex modeling of these entities.

RepLab was an initiative promoted by the EU project LiMoSINe, which aimed at studying reputation management as a “living lab”: a series of evaluation campaigns in which task design and evaluation methodologies are jointly carried out by researchers and the target user communities (reputation management experts). Given the novelty of the topic (as compared with opinion mining on product reviews and mainstream topic tracking), it was felt that an evaluation campaign would maximize the use of the data collections built within LiMoSINe, encourage the academic interest in tasks with practical relevance, and promote the standardization of evaluation methodologies and practices in the field. RepLab, therefore, set out to bring together the Information Access research community with representatives from the ORM industry, aiming at: establishing a roadmap that included a description of the language technologies required in terms of resources, algorithms, and applications; specifying suitable evaluation methodologies and metrics; developing test collections that enable systematic comparison of algorithms and reliable benchmarking of commercial systems (Amigó et al. 2012).

The activities of RepLab are described in a chapter by Carrillo-de-Albornoz et al. in this volume.

### 3.1.15 eHealth (2012–2019)

Medical content is becoming increasingly available electronically in a variety of forms ranging from patient records and medical dossiers, scientific publications and health-related websites to medical-related topics shared across social networks. Laypeople, clinicians and policy-makers need to be able to easily retrieve, and make sense of this content to support their decision making. Information retrieval systems have been commonly used as a means to access health information available online. However, the reliability, quality, and suitability of the information for the target audience varies greatly while high recall or coverage, that is finding all relevant information about a topic, is often as important as high precision, if not more. Furthermore, information seekers in the health domain also experience difficulties in expressing their information needs as search queries.<sup>16</sup>

The main objective of CLEF eHealth is thus to promote the development of information processing techniques that will assist the information provider and seeker to manage and retrieve electronically archived medical documents. The activities of this Lab are described in a chapter by Suominen et al. in this volume.

### 3.1.16 Biodiversity Identification and Prediction (2014–2019)

The LifeCLEF Lab aims at boosting research on the identification and prediction of living organisms in order to solve the taxonomic gap and improve our knowledge of biodiversity.

Building accurate knowledge of the identity, the geographic distribution and the evolution of living species is essential for a sustainable development of humanity as well as for biodiversity conservation. Unfortunately, such basic information is often only partially available for professional stakeholders, teachers, scientists and citizens, and is often incomplete for ecosystems that possess the highest diversity. A noticeable consequence of this sparse knowledge is that the precise identification of living plants or animals is usually impossible for the general public, and often difficult for professionals, such as farmers, fish farmers or foresters and even also for the naturalists and specialists themselves. This taxonomic impediment was actually identified as one of the main ecological challenges to be solved during the United Nations Conference in Rio de Janeiro in 1992. In this context, an ultimate ambition is to set up innovative information systems relying on the automated identification and understanding of living organisms as a means to engage massive crowds of observers and boost the production of biodiversity and agro-biodiversity data.<sup>17</sup>

Through its biodiversity informatics related challenges, LifeCLEF aims at pushing the boundaries of the state-of-the-art in several research directions at the frontier of multimedia information retrieval, machine learning and knowledge engineering

---

<sup>16</sup>See <https://sites.google.com/view/clef-ehealth-2018/home>.

<sup>17</sup>See <https://www.imageclef.org/lifeclef2019>.

with a focus on species identification using images for plants, audio for birds, and video for fishes.

In 2019 the LifeCLEF Lab proposes three data-oriented challenges related to this vision, in continuity with previous editions of the Lab:

- PlantCLEF aims at evaluating image-based plant identification on 10K species;
- BirdCLEF aims at evaluating bird species detection in audio soundscapes;
- GeoLifeCLEF aims at evaluating location-based prediction of species based on environmental and occurrence data.

The chapter by Joly et al. in this volume describes the activities of LifeCLEF.

### 3.1.17 News Recommendation Evaluation (2014–2017)

The NewsREEL Lab at CLEF provided the opportunity to evaluate algorithms both based on live data and offline simulated streams. The development of recommender services based on stream data is a challenging task. Systems optimized for handling streams must ensure highly precise recommendations taking into account the continuous changes in the stream as well as changes in the user preferences. In addition the technical complexity of the algorithms must be considered ensuring the seamless integration of recommendations into existing applications as well as ensuring the scalability of the system. Researchers in academia often focus on the development of algorithms only tested using static datasets due to the lack of access to live data. The benchmarking of the algorithms in the NewsREEL Lab considered both the recommendation precision (measured by the ClickThrough-Rate) and technical aspects (measured by reliability and response time) (Lommatzsch et al. 2017).

The chapter by Hopfgartner et al. in this volume includes a description of the activities of NewsReel.

### 3.1.18 Living Labs (2015–2016)

In recent years, a new evaluation paradigm known as *Living Labs* has been proposed. The idea is to perform experiments in situ, with real users doing real tasks using real-world applications. Previously, this type of evaluation had only been available to (large) industrial research labs. The main goal with the Living Labs for IR Evaluation (LL4IR) Lab at CLEF was to provide a benchmarking platform for researchers to evaluate their ranking systems in a live setting with real users in their natural task environments. The Lab acted as a proxy between commercial organizations (live environments) and lab participants (experimental systems), facilitated data exchange, and made comparison between the participating systems. This initiative was a first of its kind for IR. It dealt with evaluation of ranking systems in a live setting with real users in their natural task environments.

The chapter by Hopfgartner et al. in this volume details the activities of Living Labs.



### **3.1.19 Social Book Search (2015–2016)**

The goal of the Social Book Search (SBS) Lab was to evaluate approaches to support users in searching collections of books. The SBS Lab investigated the complex nature of relevance in book search and the role of traditional and user generated book metadata in retrieval. The aims were (1) to develop test collections to evaluate systems in terms of ranking search results and (2) to develop user interfaces and conduct user studies to investigate book search in scenarios with complex information needs and book descriptions that combine heterogeneous information from multiple sources. Techniques were studied to support users in complex book search tasks that involved more than just a query and results list, relying on semi-structured and highly structured data. The Lab included an interactive task which was a result of a merge of the INEX Social Book Search track and the Interactive task of CHiC. User interaction in social book search was gauged by observing user activity with a large collection of rich book descriptions under controlled and simulated conditions, aiming for as much “real-life” experiences as possible intruding into the experimentation. The aim was to augment the other Social Book Search tracks with a user-focused methodology. This Lab is discussed in the chapter by Kamps et al. in this volume.

### **3.1.20 Microblog Cultural Contextualization (2016–2018)**

The MC2 lab mainly focused on developing processing methods and resources to mine the social media sphere and microblogs surrounding cultural events such as festivals, concerts, books, movies and museums, dealing with languages, dialects and informal expressions. The underlying scientific problems concern both IR and the Humanities.

The Lab began with a pilot activity in 2016. This examined the contextualization of data collected on the Web, and the search of content captured or produced by internet users. Participants were given access to a massive collection of microblogs and related urls to work with. The MC2 Lab at CLEF 2017 dealt with how the cultural context of a microblog affects its social impact at large. This involved microblog search, classification, filtering, language recognition, localization, entity extraction, linking open data, and summarization. Participants had access to the massive multilingual microblog stream of The Festival Galleries project. Microblog search topics were in four languages: Arabic, English, French and Spanish, and results were expected in any language.

In the 2018 Lab, two main tasks were offered: cross-language cultural microblog search; and argumentation mining. The first task was specific to movies. Topics were extracted from the French VodKaster website that allows readers to get personal short comments (microcritics) about movies. The challenge was to find related microblogs in four different languages in a large archive. The second task was about argumentation mining, a new problem in corpus-based text analysis that addresses the challenging task of automatically identifying the justifications provided by

opinion holders for their judgment. The idea was to perform a search process on a massive microblog collection that focused on claims about a given festival. More details can be found in the chapter by Kamps et al. in this volume.

### **3.1.21 Dynamic Search for Complex Tasks (2017–2018)**

DynSe, the CLEF Dynamic Tasks Lab, attempted to focus attention towards building a bridge between batch TREC-style evaluation methodology and Interactive Information Retrieval evaluation methodology—so that dynamic search algorithms can be evaluated using reusable test collections.

Information Retrieval research has traditionally focused on serving the best results for a single query—so-called ad-hoc retrieval. However, users typically search iteratively, refining and reformulating their queries during a session. IR systems can respond to each query in a session independently of the history of user interactions, or alternatively adopt their model of relevance in the context of these interactions. A key challenge in the study of algorithms and models that dynamically adapt their response to a user's query on the basis of prior interactions is the creation of suitable evaluation resources and the definition of suitable evaluation metrics to assess the effectiveness of such IR algorithms. Over the years, various initiatives have been proposed which have tried to make progress on this long standing challenge. However, while significant effort has been made to render the simulated data as realistic as possible, generating realistic user simulation models remains an open problem (Kanoulas and Azzopardi 2017).

In its first edition, the Dynamic Search lab ran in the form of a workshop with the goal of addressing one key question: how can we evaluate dynamic search algorithms, commonly used by personalized session search, contextual search, and dialog systems. The workshop provided an opportunity for researchers to discuss the challenges faced when trying to measure and evaluate the performance of dynamic search algorithms, given the context of available corpora, simulation methods, and current evaluation metrics. To seed the discussion, a pilot task was run with the goal of producing search agents that could simulate the process of a user, interacting with a search system over the course of a search session. The outcomes of the workshop were used to define the tasks of the 2018 Lab.

### **3.1.22 Early Risk Prediction on the Internet (eRisk, 2017–2019)**

This Lab is exploring evaluation methodologies and effectiveness metrics for early risk detection on the Internet (in particular risks related to health and safety). The challenge consists of sequentially processing pieces of evidence from social media and microblogs and detecting, as soon as possible, early traces of diseases, such as depression or anorexia. For instance, early alerts could be sent when a predator starts interacting with a child for sexual purposes, or when a potential offender starts publishing antisocial threats on a blog, forum or social network. The main

goal is to pioneer a new interdisciplinary research area, potentially applicable to a wide variety of situations and to many different personal profiles. Examples include potential paedophiles, stalkers, individuals that could fall into the hands of criminal organisations, people with suicidal inclinations, or people susceptible to depression.

### 3.1.23 Evaluation of Personalised Information Retrieval (2017–2019)

The objective of the PIR-CLEF Lab is to develop and demonstrate the effectiveness of a methodology for the repeatable evaluation of Personalised Information Retrieval (PIR). PIR systems are aimed at enhancing traditional IR systems to better satisfy the information needs of individual users by providing search results that are not only relevant to the query but also to the specific user who submitted the query. In order to provide a personalised service, a PIR system maintains information about the user and their preferences and interests. These personal preferences and interests are typically inferred through a variety of interactions modes between the user with the system. This information is then represented in a user model, which is used to either improve the user's query or to re-rank a set of retrieved results so that documents that are more relevant to the user are presented in the top positions of the ranked list. Existing work on the evaluation of PIR has generally relied on a user-centered approach, mostly based on user studies; this approach involves real users undertaking search tasks in a supervised environment. While this methodology has the advantage of enabling the detailed study of the activities of real users, it has the significant drawback of not being easily reproducible and does not support the extensive exploration of the design and construction of user models and their exploitation in the search process. These limitations greatly restrict the scope for algorithmic exploration in PIR. This means that it is generally not possible to make definitive statements about the effectiveness or suitability of individual PIR methods and meaningful comparison between alternative approaches (Pasi et al. 2017).

The PIR-CLEF Lab began with a pilot task in 2017. This was undertaken by 10 users employing a clearly defined and novel methodology. Data was gathered on the activities of each participant during search sessions on a subset of the ClueWeb12 collection,<sup>18</sup> including details of relevant documents as marked by the searchers. The intention was to allow research groups working on PIR to gain experience with and provide feedback on the proposed PIR evaluation methodology. The input from the pilot task was used in the definition of the methodology employed in the 2018 and 2019 Labs. The Labs provide a framework for the evaluation of *Personalized Information Retrieval (PIR)*: (1) to facilitate comparative evaluation by offering participating research groups a mechanism for the evaluation of their personalisation algorithms; (2) to give the participating groups the means to formally define and evaluate their own novel user profiling approaches for PIR.

---

<sup>18</sup><https://lemurproject.org/clueweb12/>.

This is the first evaluation benchmark based on the Cranfield paradigm in this research area, with the potential benefits of producing evaluation results that are easily reproducible.

### **3.1.24 Automatic Identification and Verification of Political Claims (2018–2019)**

The CheckThat! Lab aims at fostering the development of technology capable of both spotting and verifying check-worthy claims in political debates in English and Arabic. Investigative journalists and volunteers work hard trying to get to the root of a claim in order to present solid evidence in favor or against it. However, manual fact-checking is very time-consuming, and automatic methods have been proposed as a way of speeding-up the process. For instance, there has been work on checking the factuality/credibility of a claim, of a news article, or of an entire news outlet. However, less attention has been paid to other steps of the fact-checking pipeline, e.g., check worthiness estimation has been severely understudied as a problem. By comparing a claim against the retrieved evidence, a system can determine whether the claim is likely true or likely false (or unsure, if no supporting evidence either way could be found). CheckThat! aims to address these understudied aspects. It is fostering the development of technology capable of spotting check-worthy claims in English political debates in addition to providing evidence-supported verification of Arabic claims.

### **3.1.25 Reproducibility (2018–2019)**

The goal of CENTRE@CLEF is to run a joint task across CLEF/NTCIR/TREC on reproducibility, a primary concern in many areas of science.

Information Retrieval is especially interested in reproducibility since it is a discipline strongly rooted in experimentation, where experimental evaluation represents a main driver of advancement and innovation. In 2015, the ECIR conference began a new track focused on the reproducibility of previously published results. This conference track led to 3–4 reproducibility papers accepted each year but, unfortunately, this valuable effort did not produce a systematic approach to reproducibility: submitting authors adopted different notions of reproducibility, they adopted very diverse experimental protocols, they investigated the most disparate topics, resulting in a very fragmented picture of what was reproducible and what not, and the results of these reproducibility papers are spread over a series of potentially disappearing repositories and Web sites. It is clear that there is a need and urgency for a systematic approach to reproducibility in IR. The joint task at CENTRE@CLEF challenges participants:

- to reproduce the best results of the best/most interesting systems in previous editions of CLEF/NTCIR/TREC by using standard open source IR systems;

- to provide the community with the additional components and resources that were developed to reproduce the results with the hope of improving existing open source systems.

## 4 IR Tools and Test Collections

CLEF activities over these last two decades have resulted in the creation of a considerable amount of valuable resources, extremely useful for many types of text processing and benchmarking activities in the IR domain. In this section, we provide some pointers with respect to their availability.

Much attention was paid in the first years of CLEF 1.0 to the processing requirements of different languages; these vary considerably depending on levels of morphological and syntactic complexity. This resulted in many comparative studies and the development of a variety of morphological processors (light and more aggressive stemmers), see the discussion in the chapter by Savoy and Braschler in this volume. Jacques Savoy also maintains an important site at the University of Neuchâtel which provides information on and links to many IR multilingual tools.<sup>19</sup>

The test collections, created as a result of the diverse experimental evaluation initiatives conducted in CLEF represent the end results of much collaborative work aimed at providing understanding and insights into how system performances can best be improved and how progress can be achieved. As already stated, the CLEF evaluation campaigns have mainly adopted a comparative evaluation approach in which system performances are compared according to the Cranfield methodology (see the chapter by Voorhees for a description of Cranfield). The test collections produced are thus made up of documents, topics and relevance assessments. The topics are created to simulate particular information needs from which the systems derive the queries to search the document collections. System performance is evaluated by judging the results retrieved in response to a topic with respect to their relevance, and computing the relevant measures, depending on the methodology adopted by the Track/Lab. The chapter by Agosti et al. in this volume describes the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* system which manages and provides access to much of the data used and produced within CLEF.

During the campaigns, participating groups are provided with access to the necessary data sets on signing a data agreement form which specifies the conditions of use. An objective of CLEF is that, at the end of an evaluation, the test collections produced should, whenever possible, be made available to the wider R&D community. Here below we give some examples of collections that are now publicly accessible. If you do not find what you were looking for, our advice is to

---

<sup>19</sup><http://members.unine.ch/jacques.savoy/clef/>.

contact the coordinators of the relevant Track or Lab to see if they can help you. Contact information can be found via the CLEF web site<sup>20</sup> and/or annual working notes.<sup>21</sup>

## 4.1 *ELRA Catalogue*

A number of official CLEF Test Suites consisting of the data created for the monolingual, bilingual, multilingual and domain-specific text retrieval and question answering tracks in the CLEF 1.0 Campaigns are available, generally for a fee, in the catalogue of the European Language Resources Association (ELRA).<sup>22</sup> These packages consist of multilingual document collections in many languages; step-by-step documentation on how to perform system evaluation; tools for results computation; multilingual sets of topics; multilingual sets of relevance assessments; guidelines for participants (in English); tables of the results obtained by the participants; publications. The following data collections are included:

- CLEF multilingual corpus of more than 3 million news documents in 14 European languages. This corpus is divided into two comparable collections: 1994–1995—Dutch, English, Finnish, French, German, Italian, Portuguese, Russian, Spanish, Swedish; 2000–2002—Basque, Bulgarian, Czech, English, Hungarian. These collections were used in the Ad-Hoc, and Question Answering packages.
- The GIRT-4 social science database in English and German (over 300,000 documents) and two Russian databases: the Russian Social Science Corpus (approx. 95,000 documents) and the Russian ISSS collection for sociology and economics (approx. 150,000 docs); Cambridge Sociological Abstracts in English (20,000 docs). These collections were used in the domain-specific package.

The ELRA catalog also lists test suites derived from CLEF eHealth activities. These packages contain data used for user-centred health information retrieval tasks conducted at the CLEF eHealth Labs in 2013 and 2014 and include: a collection of medical-related documents in English; guidelines provided to the participants; queries generated by medical professionals in several languages; a set of manual relevance assessments; the official results obtained by the participants; working notes papers.

---

<sup>20</sup><http://www.clef-initiative.eu/>.

<sup>21</sup>In the CEUR Workshop Proceedings—<http://http://ceur-ws.org/>.

<sup>22</sup>Information and conditions of purchase can be found at: <http://catalog.elra.info/>.

## 4.2 *Some Publicly Accessible CLEF Test Suites*

Many Labs make evaluation test suites available free-of-charge for research and system training purposes. Here below, we list what is currently available at the time of writing (April 2019).

- QA@CLEF: Question Answering  
In addition to what can be found on the ELRA Catalogue, datasets for advanced tasks are accessible at <http://nlp.uned.es/clef-qa/repository/pastCampaigns.php>
- PAN: Digital Text Forensics  
Datasets designed for Authorship, Author Profiling, Credibility Analysis, Deception Detection, and Text Reuse Detection tasks. Accessible at <https://pan.webis.de/data.html>.
- RepLab: Online Reputation Management
  - RepLab 2013: +500,000 reputation expert annotations on Twitter data, covering named entity disambiguation (filtering task), reputational polarity, topic detection and topic reputational priority (alert detection). Accessible at <http://nlp.uned.es/replab2013/>
  - RepLab 2014: additional annotations on RepLab 2013 tweets covering reputational dimensions of tweets (Products/Services, Innovation, Workplace, Citizenship, Governance, Leadership, and Performance) and author profiling: (1) identification of opinion makers and (2) classification of author types (journalist, professional, authority, activist, investor, company or celebrity). Accessible at <http://nlp.uned.es/replab2014/>
- WePS: Web People Search  
WePS 3 included two tasks concerning the Web entity search problem:
  - Task 1 is related to Web People Search and focuses on person name ambiguity and person attribute extraction on Web pages;
  - Task 2 is related to Online Reputation Management (ORM) for organizations and focuses on the problem of ambiguity for organization names and the relevance of Web data for reputation management purposes. Test collections accessible at <http://nlp.uned.es/weps/weps-3>

Previous WePS datasets are also accessible at <http://nlp.uned.es/weps/weps-1/weps1-data> and <http://nlp.uned.es/weps/weps-2>
- Social Book Search  
2.8 million book records in XML format. Accessible at <http://social-book-search.humanities.uva.nl/>
- Protest News
  - Annotated data from the publicly available English Reuters news text Corpus RCV1 will be made freely accessible. See Lab website for details.

- The ImageCLEF and LifeCLEF initiatives make a number of existing datasets for system training purposes. Full details and information concerning conditions of use of the following collections can be found at the ImageCLEF website, see <https://www.imageclef.org/datasets>.
  - ImageCLEF/IAPR TC 12 Photo Collection
  - Segmented IAPR dataset
  - The COLD Database: contains image sequences captured using a regular and omni-directional cameras mounted on different mobile robot platforms together with laser range scans and odometry data. Data recorded at three different indoor laboratory environments located in three different European cities under various weather and illumination conditions.
  - The IDOL2 Database: consists of 24 image sequences accompanied by laser scans and odometry data acquired using two mobile robot platforms, within an indoor laboratory environment consisting of five rooms of different functionality, under various illumination conditions and across a span of 6 months.
  - The INDECS Database: several sets of pictures taken indoors, in five rooms of different functionality under various illumination and weather conditions at different periods of time.
  - ImageCLEF VCDT test collections: test collections of the ImageCLEF Visual Concept Detection and Annotation Task (VCDT) from 2009–2011
  - ImageCLEF Wikipedia Image Retrieval Datasets—The Wikipedia image retrieval task ran as part of ImageCLEF for 4 years: 2008–2011.
- Other test collections used in ImageCLEF tasks are listed here:
  - 2012 ImageCLEF WEBUPV Collection: images crawled from the web and web pages that contained them. 253000 images. Accessible at <http://doi.org/10.5281/zenodo.1038533>
  - 2013 ImageCLEF WEBUPV Collection: images crawled from the web and web pages that contained them. 253000 images. Accessible at <http://doi.org/10.5281/zenodo.257722>
  - 2014 ImageCLEF WEBUPV Collection: images crawled from the web and web pages that contained them. 505122 images. Accessible at <http://doi.org/10.5281/zenodo.259758>
  - 2015 ImageCLEF WEBUPV Collection: images crawled from the web and web pages that contained them. 500000 images. Accessible at <http://doi.org/10.5281/zenodo.1038547>
  - 2016 ImageCLEF WEBUPV Collection: images crawled from the web and web pages that contained them. 500000 images. Accessible at <http://doi.org/10.5281/zenodo.1038554>
  - ImageCLEF 2016 Bentham Handwritten Retrieval Dataset: images of scanned pages of a manuscript and queries to retrieve. Language: English. Size: 363 pages train, 433 pages development, 200 pages test. Accessible at <http://doi.org/10.5281/zenodo.52994>



## 5 The CLEF Association

The CLEF Association<sup>23</sup> is an independent non-profit legal entity, established in October 2013 as a result of the activity of the PROMISE Network of Excellence,<sup>24</sup> which sponsored CLEF from 2010 to 2013.

The Association has scientific, cultural and educational objectives and operates in the field of information access systems and their evaluation. Its mission is:

- to promote access to information and use evaluation;
- to foster critical thinking about advancing information access and use from a technical, economic and societal perspective.

Within these two areas of interest, the CLEF Association aims at a better understanding of the use and access to information and how to improve this. The two areas of interest translate into the following objectives:

- providing a forum for stakeholders with multidisciplinary competences and different needs, including academia, industry, education and other societal institutions;
- facilitating medium/long-term research in information access and use and its evaluation; increasing, transferring and applying expertise.

The CLEF Association currently plays a key role in CLEF by ensuring the continuity, self-sustainability and overall coordination. CLEF 2014 was the first edition of CLEF not supported in any way by a main European project but run on a totally volunteer basis with the support of the CLEF association membership fees paid by its multidisciplinary research community.

## 6 Impact

Shared evaluation campaigns have always played a central role in IR research. They have produced huge improvements in the state-of-the-art and helped solidify a common systematic methodology, achieving not only scholarly impact (Tsirikla et al. 2013, 2011; Thornley et al. 2011; Angelini et al. 2014) but also economic results (Rowe et al. 2010), estimated in a return-on-investment about 3–5 times the funding provided. The 20 years of CLEF campaigns have had a significant scientific impact on European and global research. This is documented in the chapter by Birger Larsen in the final part of this volume.

During their life-span, these large-scale campaigns also produce a huge amount of extremely valuable experimental data. This data provides the foundations for sub-

---

<sup>23</sup><http://www.clef-initiative.eu/association>.

<sup>24</sup><http://www.promise-noe.eu/>.

sequent scientific production and system development and constitutes an essential reference for the literature in the field. Papers by Agosti et al., Müller and Hanbury, and Potthast et al. in this volume explore the infrastructures developed in CLEF over the years to run the experiments and to manage the resulting experimental data. Section 4 provides information on the availability of many of the IR resources and test collections created as a result of CLEF experiments.

Up until the end of the Twentieth century, IR research was predominantly conducted on test collections in English. Thus, when we launched CLEF 1.0, one of our declared objectives was to stimulate research in our domain on collections in many different languages—not only English—and across language boundaries. As a European initiative our primary focus was on European languages. This is the topic of the chapter by Savoy and Braschler. This goal was so well achieved that in CLEF 2.0 we could almost state that *multilinguality* in European IR research activities is taken for granted; even if the main theme is *multimodality*, all of the CLEF 2.0 Labs handle data in more than one European language.

Another of our goals has been to impact not only academia but also industrial research. IR research can never be considered only at the theoretical level, clearly the over-riding factors are the requirements of society at large. An important step in this direction, which began in CLEF 1.0 with ImageCLEF medical retrieval experiments (see the chapter by Müller et al. in this volume) but has certainly been increasingly reinforced in CLEF 2.0, is the involvement of real world user communities. Thus, just to cite a few examples, we have seen collaborations with the intellectual property and patent search domain in CLEF-IP (see the chapter by Piroi and Hanbury), with health specialists in E-Health (Suominen et al. this volume), and with news portals in the NewsREEL project (see Hopfgartner et al.). The chapter by Jussi Karlgren in the final part of this volume discusses the challenges involved in applying evaluation benchmarks in operational settings. And this year, CLEF 2019 will host for the first time an Industry Day, jointly organized with the Swiss Alliance for Data-Intensive Services. The goal is to further open CLEF to a wider, industrial community through demo sessions, panels and special keynotes where the very best and most pertinent work of CLEF participants will be made more publicly accessible.

An aspect of CLEF of which we are particularly proud is the consolidation of a strong community of European researchers in the multidisciplinary context of IR. This year, for the first time, the *European Conference for Information Retrieval (ECIR)* and CLEF have joined forces: ECIR 2019 hosting a session dedicated to CLEF Labs where lab organizers present the major outcomes of their Labs and plans for ongoing activities, followed by a poster session in order to favour discussion during the conference. This is reflected in the ECIR 2019 proceedings, where CLEF Lab activities and results are reported as short papers. The goal is not only to engage the ECIR community in CLEF activities, but also to disseminate the research results achieved during CLEF evaluation cycles at ECIR. This collaboration will of course strengthen European IR research even more. However, this European community should not be seen in isolation. CLEF is part of a global community; we have always maintained close links with our peer initiatives in the Americas and Asia. There is a

strong bond connecting TREC, NTCIR, CLEF and FIRE, and a continual, mutually beneficial exchange of ideas, experiences and results.

Despite the acknowledged success of CLEF and other evaluation campaigns over the years, we cannot rest on our laurels. It is fundamental to keep asking what new challenges need to be addressed in the future and how to continue to contribute to progress in the IR field. The chapters in the concluding part of this volume thus explore future perspectives: reproducibility of experiments by Norbert Fuhr, industrial involvement by Jussi Karlgren, and exploitation of Visual Analytics for IR evaluation by Ferro and Santucci.

**Acknowledgements** CLEF 2000 and 2001 were supported by the European Commission under the Information Society Technologies programme and within the framework of the DELOS Network of Excellence for Digital Libraries (contract no. IST-1999-12262).

CLEF 2002 and 2003 were funded as an independent project (contract no. IST-2000-31002) under the 5th Framework Programme of the European Commission.

CLEF 2004–2007 were sponsored by the DELOS Network of Excellence for Digital Libraries (contract no. G038-507618) under the 6th Framework Programme of the European Commission.

Under the 7th Framework Programme of the European Commission, CLEF 2008 and 2009 were supported by TrebleCLEF Coordination Action (contract no. 215231) and CLEF 2010 to 2013 were funded by the PROMISE Network of Excellence (contract no. 258191).

CLEF 2011–2015 also received support from the ELIAS network (contract no. 09-RNP-085) of the European Science Foundation (ESF) for ensuring student travel grants and invites speakers.

CLEF 2015, 2017, and 2018 received ACM SIGIR support for student travel grants through the SIGIR Friends program.

Over the years CLEF has also attracted industrial sponsorship: from 2010 onwards, CLEF has received the support of Google, Microsoft, Yandex, Xerox, Celi as well as publishers in the field such as Springer and Now Publishers.

In addition to the support gratefully acknowledged above, CLEF tracks and labs have frequently received the assistance of other projects and organisations; unfortunately, it is impossible to list them all here.

It must be noted that, above all, CLEF would not be possible without the volunteer efforts, enthusiasm, and passion of its community: lab organizers, lab participants, and attendees are the core and the real success of CLEF.

## References

- Amigó E, Corujo A, Gonzalo J, Meij E, de Rijke M (2012) Overview of RepLab 2012: evaluating online reputation management systems. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Amigó E, Gonzalo J, Verdejo MF (2013) A general evaluation measure for document organization tasks. In: Jones GJF, Sheridan P, Kelly D, de Rijke M, Sakai T (eds) Proc. 36th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2013). ACM Press, New York, pp 643–652
- Angelini M, Ferro N, Larsen B, Müller H, Santucci G, Silvello G, Tsikrika T (2014) Measuring and analyzing the scholarly impact of experimental evaluation initiatives. In: Agosti M, Catarci T, Esposito F (eds) Proc. 10th Italian research conference on digital libraries (IRCDL 2014). *Procedia computer science*, vol. 38, pp 133–137

- Angelini M, Fazzini V, Ferro N, Santucci G, Silvello G (2018) CLAIRE: a combinatorial visual analytics system for information retrieval evaluation. *Inf Process Manag* 54(6):1077–1100
- Bollmann P (1984) Two axioms for evaluation measures in information retrieval. In: van Rijsbergen CJ (ed) *Proc. of the third joint BCS and ACM symposium on research and development in information retrieval*. Cambridge University Press, Cambridge, pp 233–245
- Braschler M (2004) Combination approaches for multilingual text retrieval. *Inf Retr* 7(1/2):183–204
- Buckley C, Voorhees EM (2004) Retrieval evaluation with incomplete information. In: Sanderson M, Järvelin K, Allan J, Bruza P (eds) *Proc. 27th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2004)*. ACM Press, New York, pp 25–32
- Buckley C, Voorhees EM (2005) Retrieval system evaluation. In: Harman DK, Voorhees EM (eds) *TREC. Experiment and evaluation in information retrieval*. MIT Press, Cambridge, pp 53–78
- Busin L, Mizzaro S (2013) Axiometrics: an axiomatic approach to information retrieval effectiveness metrics. In: Kurland O, Metzler D, Lioma C, Larsen B, Ingwersen P (eds) *Proc. 4th international conference on the theory of information retrieval (ICTIR 2013)*. ACM Press, New York, pp 22–29
- Büttcher S, Clarke CLA, Yeung PCK, Soboroff I (2007) Reliable information retrieval evaluation with incomplete and biased judgements. In: Kraaij W, de Vries AP, Clarke CLA, Fuhr N, Kando N (eds) *Proc. 30th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2007)*. ACM Press, New York, pp 63–70
- Carterette BA (2012) Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Trans Inf Syst* 30(1):4:1–4:34
- Chapelle O, Metzler D, Zhang Y, Grinspan P (2009) Expected reciprocal rank for graded relevance. In: Cheung DWL, Song IY, Chu WW, Hu X, Lin JJ (eds) *Proc. 18th international conference on information and knowledge management (CIKM 2009)*. ACM Press, New York, pp 621–630
- Cleverdon CW (1967) The cranfield tests on index languages devices. *ASLIB Proc* 19(6):173–194
- Di Nunzio GM, Leveling J, Mandl T (2011) LogCLEF 2011 multilingual log file analysis: language identification, query classification, and success of a query. In: Petras V, Forner P, Clough P, Ferro N (eds) *CLEF 2011 working notes*. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1177/>
- Ferrante M, Ferro N, Maistro M (2015) Towards a formal framework for utility-oriented measurements of retrieval effectiveness. In: Allan J, Croft WB, de Vries AP, Zhai C, Fuhr N, Zhang Y (eds) *Proc. 1st ACM SIGIR international conference on the theory of information retrieval (ICTIR 2015)*. ACM Press, New York, pp 21–30
- Ferrante M, Ferro N, Pontarollo S (2017) Are IR evaluation measures on an interval scale? In: Kamps J, Kanoulas E, de Rijke M, Fang H, Yilmaz E (eds) *Proc. 3rd ACM SIGIR international conference on the theory of information retrieval (ICTIR 2017)*. ACM Press, New York, pp 67–74
- Ferrante M, Ferro N, Pontarollo S (2019) A general theory of IR evaluation measures. *IEEE Trans Knowl Data Eng* 31(3):409–422
- Ferro N, Harman D (2010) CLEF 2009: Grid@CLEF pilot track overview. In: Peters C, Di Nunzio GM, Kurimo M, Mandl T, Mostefa D, Peñas A, Roda G (eds) *Multilingual information access evaluation, vol. I text retrieval experiments – tenth workshop of the cross-language evaluation forum (CLEF 2009)*. Revised selected papers. Lecture notes in computer science (LNCS), vol 6241. Springer, Heidelberg, pp 552–565
- Ferro N, Silvello G (2016) A general linear mixed models approach to study system component effects. In: Perego R, Sebastiani F, Aslam J, Ruthven I, Zobel J (eds) *Proc. 39th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2016)*. ACM Press, New York, pp 25–34
- Ferro N, Silvello G (2017) Towards an anatomy of IR system component performances. *J Am Soc Inf Sci Technol* 69(2):187–200
- Ferro N, Maistro M, Sakai T, Soboroff I (2018) Overview of CENTRE@CLEF 2018: a first tale in the systematic reproducibility realm. In: Bellot P, Trabelsi C, Mothe J, Murtagh F, Nie JY,

- Soulier L, SanJuan E, Cappellato L, Ferro N (eds) Experimental IR meets multilinguality, multimodality, and interaction. Proceedings of the ninth international conference of the CLEF association (CLEF 2018). Lecture notes in computer science (LNCS), vol 11,018. Springer, Heidelberg, pp 239–246
- Fuhr N (2012) Salton award lecture: information retrieval as engineering science. *SIGIR Forum* 46(2):19–28
- Hanbury A, Müller H (2010) Automated component-level evaluation: present and future. In: Agosti M, Ferro N, Peters C, de Rijke M, Smeaton A (eds) Multilingual and multimodal information access evaluation. Proceedings of the international conference of the cross-language evaluation forum (CLEF 2010). Lecture notes in computer science (LNCS), vol 6360. Springer, Heidelberg, pp 124–135
- Harman DK (2011) Information retrieval evaluation. Morgan & Claypool Publishers, San Rafael
- Harman DK, Voorhees EM (eds) (2005) TREC. Experiment and evaluation in information retrieval. MIT Press, Cambridge
- Harman DK, Braschler M, Hess M, Kluck M, Peters C, Schäuble P, Sheridan P (2001) CLIR evaluation at TREC. In: Peters C (ed) Cross-language information retrieval and evaluation: workshop of cross-language evaluation forum (CLEF 2000). Lecture notes in computer science (LNCS), vol 2069. Springer, Heidelberg, pp 7–23
- Hopfgartner F, Hanbury A, Müller H, Eggel I, Balog K, Brodt T, Cormack GV, Lin J, Kalpathy-Cramer J, Kando N, Kato MP, Krithara A, Gollub T, Potthast M, Viegas E, Mercer S (2018) Evaluation-as-a-service for the computational sciences: overview and outlook. *ACM J Data Inf Qual* 10(4):15:1–15:32
- Hull DA (1993) Using statistical testing in the evaluation of retrieval experiments. In: Korfhage R, Rasmussen E, Willett P (eds) Proc. 16th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 1993). ACM Press, New York, pp 329–338
- Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of IR techniques. *ACM Trans Inf Syst* 20(4):422–446
- Kanoulas E, Azzopardi L (2017) CLEF 2017 dynamic search evaluation lab overview. In: Jones GJF, Lawless S, Gonzalo J, Kelly L, Goeuriot L, Mandl T, Cappellato L, Ferro N (eds) Experimental IR meets multilinguality, multimodality, and interaction. Proceedings of the eighth international conference of the CLEF association (CLEF 2017). Lecture notes in computer science (LNCS), vol 10,456. Springer, Heidelberg, pp 361–366
- Kekäläinen J, Järvelin K (2002) Using graded relevance assessments in IR evaluation. *J Am Soc Inf Sci Technol* 53(13):1120–1129
- Kelly D (2009) Methods for evaluating interactive information retrieval systems with users. *Found Trends Inf Retr* 3(1–2):1–224
- Lommatzsch A, Kille B, Hopfgartner F, Larson M, Brodt T, Seiler J, Özgöbek Ö (2017) CLEF 2017 NewsREEL overview: a stream-based recommender task for evaluation and education. In: Jones GJF, Lawless S, Gonzalo J, Kelly L, Goeuriot L, Mandl T, Cappellato L, Ferro N (eds) Experimental IR meets multilinguality, multimodality, and interaction. Proceedings of the eighth international conference of the CLEF association (CLEF 2017). Lecture notes in computer science (LNCS), vol 10,456. Springer, Heidelberg, pp 239–254
- McNamee P, Mayfield J (2004) Character N-gram tokenization for European language text retrieval. *Inf Retr* 7(1–2):73–97
- Mizzaro S (1997) Relevance: the whole history. *J Am Soc Inf Sci Technol* 48(9):810–832
- Moffat A, Zobel J (2008) Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans Inf Syst* 27(1):2:1–2:27
- Nardi A, Peters C, Vicedo JL, Ferro N (eds) (2006) CLEF 2006 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1172/>
- Nardi A, Peters C, Ferro N (eds) (2007) CLEF 2007 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1173/>
- Pasi G, Jones GJF, Marrara S, Sanvitto C, Ganguly D, Sen P (2017) Overview of the CLEF 2017 personalised information retrieval pilot lab (PIR-CLEF 2017). In: Jones GJF, Lawless

- S, Gonzalo J, Kelly L, Goeuriot L, Mandl T, Cappellato L, Ferro N (eds) Experimental IR meets multilinguality, multimodality, and interaction. Proceedings of the eighth international conference of the CLEF association (CLEF 2017). Lecture notes in computer science (LNCS), vol 10,456. Springer, Heidelberg, pp 338–345
- Peters C (ed) (2001) Cross-language information retrieval and evaluation: workshop of cross-language evaluation forum (CLEF 2000). Lecture notes in computer science (LNCS), vol 2069. Springer, Heidelberg
- Rowe BR, Wood DW, Link AL, Simoni DA (2010) Economic impact assessment of NIST's text retrieval conference (TREC) program. RTI Project Number 0211875, RTI International. <http://trec.nist.gov/pubs/2010.economic.impact.pdf>
- Sakai T (2006) Evaluating evaluation metrics based on the bootstrap. In: Efthimiadis EN, Dumais S, Hawking D, Järvelin K (eds) Proc. 29th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2006). ACM Press, New York, pp 525–532
- Sakai T (2012) Evaluation with informational and navigational intents. In: Mille A, Gandon FL, Misselis J, Rabinovich M, Staab S (eds) Proc. 21st international conference on world wide web (WWW 2012). ACM Press, New York, pp 499–508
- Sakai T (2014a) Metrics, statistics, tests. In: Ferro N (ed) Bridging between information retrieval and databases - PROMISE winter school 2013, revised tutorial lectures. Lecture notes in computer science (LNCS), vol 8173. Springer, Heidelberg, pp 116–163
- Sakai T (2014b) Statistical reform in information retrieval? SIGIR Forum 48(1):3–12
- Sanderson M (2010) Test collection based evaluation of information retrieval systems. *Found Trends Inf Retr* 4(4):247–375
- Saracevic T (1975) RELEVANCE: a review of and a framework for the thinking on the notion in information science. *J Am Soc Inf Sci Technol* 26(6):321–343
- Savoy J (1997) Statistical inference in retrieval effectiveness evaluation. *Inf Process Manag* 33(4):495–512
- Spärck Jones K (ed) (1981) Information retrieval experiment. Butterworths, London
- Thornley CV, Johnson AC, Smeaton AF, Lee H (2011) The scholarly impact of TRECVID (2003–2009). *J Am Soc Inf Sci Technol* 62(4):613–627
- Tsikrika T, Garcia Seco de Herrera A, Müller H (2011) Assessing the scholarly impact of image CLEF. In: Forner P, Gonzalo J, Kekäläinen J, Lalmas M, de Rijke M (eds) Multilingual and multimodal information access evaluation. Proceedings of the second international conference of the cross-language evaluation forum (CLEF 2011). Lecture notes in computer science (LNCS), vol 6941. Springer, Heidelberg, pp 95–106
- Tsikrika T, Larsen B, Müller H, Endrullis S, Rahm E (2013) The scholarly impact of CLEF (2000–2009). In: Forner P, Müller H, Paredes R, Rosso P, Stein B (eds) Information access evaluation meets multilinguality, multimodality, and visualization. Proceedings of the fourth international conference of the CLEF initiative (CLEF 2013). Lecture notes in computer science (LNCS), vol 8138. Springer, Heidelberg, pp 1–12
- van Rijsbergen CJ (1974) Foundations of evaluation. *J Doc* 30(4):365–373
- Voorhees EM, Harman DK (1998) Overview of the seventh text retrieval conference (TREC-7). In: Voorhees EM, Harman DK (eds) The seventh text retrieval conference (TREC-7). National Institute of Standards and Technology (NIST), Special Publication 500-242, Washington, pp 1–24
- Yilmaz E, Aslam JA (2006) Estimating average precision with incomplete and imperfect judgments. In: Yu PS, Tsotras V, Fox EA, Liu CB (eds) Proc. 15th international conference on information and knowledge management (CIKM 2006). ACM Press, New York, pp 102–111

# The Evolution of Cranfield



Ellen M. Voorhees

**Abstract** Evaluating search system effectiveness is a foundational hallmark of information retrieval research. Doing so requires infrastructure appropriate for the task at hand, which generally follows the Cranfield paradigm: test collections and associated evaluation measures. A primary purpose of *Information Retrieval (IR)* evaluation campaigns such as *Text REtrieval Conference (TREC)* and *Conference and Labs of the Evaluation Forum (CLEF)* is to build this infrastructure. The first TREC collections targeted the same task as the original Cranfield tests and used measures that were familiar to test collection users of the time. But as evaluation tasks have multiplied and diversified, test collection construction techniques and evaluation measure definitions have also been forced to evolve. This chapter examines how the Cranfield paradigm has been adapted to meet the changing requirements for search systems enabling it to continue to support a vibrant research community.

## 1 Introduction

Information retrieval research has a rich tradition of experimentation. In the 1960s, Cyril Cleverdon and his colleagues at the College of Aeronautics, Cranfield, ran a series of tests to determine appropriate indexing languages—schemes to represent document content that would enable trained search intermediaries to find appropriate references for library patrons (Cleverdon 1967). The conclusion reached in the experiments, that a document’s own words are effective for indexing, was highly controversial at the time though generally accepted today. The experiments are best known, however, for being the first to use a test collection. By comparing the effectiveness of different languages on a common document set with a common set

---

E. M. Voorhees (✉)  
National Institute of Standards and Technology, Gaithersburg, MD, USA  
e-mail: [ellen.voorhees@nist.gov](mailto:ellen.voorhees@nist.gov)

This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2019  
N. Ferro, C. Peters (eds.), *Information Retrieval Evaluation in a Changing World*, The Information Retrieval Series 41,  
[https://doi.org/10.1007/978-3-030-22948-1\\_2](https://doi.org/10.1007/978-3-030-22948-1_2)

of information needs, Cleverdon was able to control for much of the variability that had plagued earlier attempts to compare languages and in the process established what has become known as the Cranfield paradigm.

Test collections have been used in information retrieval research since then, though they have also had their detractors from the start (Taube 1965; Cuadra and Katter 1967). Early detractors feared the fluidity of ‘relevance’ made it unsuitable to be a fundamental component of an evaluation strategy. Later, concerns arose over the unrealistically small size of test collections compared to the data sets of operational systems, as well as the use of incompatible evaluation measures by different research groups that prevented their published retrieval results from being truly comparable. By the early 1990s even some Cranfield practitioners were questioning whether test collections had out-lived their usefulness (Robertson and Hancock-Beaulieu 1992).

In response to these concerns, in 1992 the U.S. *National Institute of Standards and Technology (NIST)* founded the TREC workshop with the goal of building a single realistically-large test collection to support IR research. TREC not only accomplished that goal, but along with its companion evaluation conferences such as CLEF, *NII Testbeds and Community for Information access Research (NTCIR)*, and *Forum for Information Retrieval Evaluation (FIRE)* that followed, went on to build dozens of large test collections for a variety of tasks. In addition, the conferences have standardized and validated best practices in the use of test collections.

This chapter examines how the Cranfield paradigm has been adapted to meet the changing requirements for information retrieval research in the era of community evaluation conferences. The next section gives a short recap of Cranfield before TREC to give context; see Robertson (2008) and Sanderson (2010) for more detailed accounts of the history of IR evaluation in this period. The following section describes research surrounding the early TREC collections that both enabled larger collections to be built and validated the existing experimental protocol. Section 4 then examines some of the ways the Cranfield paradigm has been extended in support of research in other areas and modern IR research problems.

## 2 Cranfield Pre-TREC

The Cranfield paradigm can be summarized as follows. A test collection consists of a set of documents, a set of information need statements (called “topics” in the remainder of this chapter), and a set of relevance judgments that list which documents should be returned for which topic. A researcher runs a retrieval system on a test collection to produce a ranked list of documents in response to each topic (this is a “run”). A ranking reflects the system’s idea of which documents are likely to be relevant to the topic; documents it believes more likely to be relevant are ranked ahead of documents it believes are less likely to be relevant. Using the relevance judgments, some evaluation metric is computed for the ranked list for each topic and scores for individual topics are averaged over the set of topics in the



test collection. Different systems produce runs for the exact same test collection, and the average scores are compared. Retrieval systems producing runs with better average scores are considered more effective retrieval systems.

Cranfield is an intentionally stark abstraction of any real user search task, representing what Spärck Jones (2001) calls the core competency of search. The abstraction arises through three major simplifying assumptions:

- relevance can be approximated by topical similarity, which implies all relevant documents are equally desirable; relevance of one document is independent of the relevance of any other document; and the user information need is static.
- a single set of judgments for a topic is representative of the user population.
- (essentially) all relevant documents for a topic are known.

While these assumptions are not true, the abstraction represents a fundamental capability that any actual search system must possess. It is hard to imagine how a search system could be effective if it cannot at least distinguish relevant documents from not relevant ones.

Use of the Cranfield methodology also requires an evaluation measure. Early work in information retrieval, including the Cranfield tests, produced retrieved sets of documents (as opposed to ranked lists), and measured the effectiveness of retrieval in terms of the *precision* and *recall* of the set (Keen 1966). Precision is the fraction of retrieved documents that are relevant, and recall the fraction of relevant documents that are retrieved. In practice, the two measures vary inversely with one another, so both measures were needed to get an accurate view of the quality of the system. The advent of using ranked output required adapting the measurement methodology to define the retrieved set. This was generally done by defining a cut-off level, the rank such that everything in the list at or before that rank was considered retrieved and everything after it not retrieved. Alternatively, precision could be reported for a standard set of recall values, for example precision at 0.25, 0.5, and 0.75 recall. The use of standard recall values requires interpolation since the actual recall values that are obtainable for a given topic depends on the number of relevant documents the topic has, and there are various methods by which the interpolation could be performed. Averaging the results over a set of topics also has different options. Using precision as an example, one can compute the precision on a per-topic basis and then take the mean over the set of topics, or one can divide the total number of relevant documents retrieved for all topics by the total number of documents retrieved for all topics. Averaging schemes, and especially interpolation schemes, were the subject of much debate in these early years.

During the 25 years following the Cranfield experiments, other retrieval test collections were built and these collections were often shared among different research groups. But there was no agreement on which measures to use and measures proliferated. Research papers of the time generally reported only the authors' own favorite measures, and even when two papers reported values for what was called the same measure the actual implementations of that measure differed (interpolation differences), leading to incomparable results. Research groups could not build on one another's work because there was no common basis to do so.

The test collections in use by the research community were also small in comparison to the document set sizes used in commercial retrieval systems. Commercial retrieval systems were searching document sets that were orders of magnitude larger than the publicly available test sets of the time, and it was believed that operators of the commercial systems would continue to discount research results unless those results were demonstrated on comparably-sized collections (Ledwith 1992). Cleverdon had made the deliberate decision to use small document set sizes so that all documents could be judged for all topics, known as complete judgments (Cleverdon 1991). Unfortunately, even for document set sizes of several thousand documents, getting complete judgments is an arduous task; getting them for very much larger collections is out of the question.

In the latter half of the 1970s, Karen Spärck Jones and colleagues argued the need for and proposed how to build a large (30,000 documents!), general-purpose test collection that they called the 'IDEAL' collection (Spärck Jones and Van Rijsbergen 1975; Spärck Jones and Bates 1977; Gilbert and Spärck Jones 1979). The desire for a general-purpose collection acknowledged the problem that the collections being shared had each been developed to test a specific hypothesis and were not necessarily appropriate for the different research questions being investigated in subsequent use. The proposal was wide-ranging, touching on many different aspects of test collection methodology, but in particular suggested using pooling to obtain essentially complete judgments without actually having all documents in the collection judged. The essential idea of pooling is to obtain a human judgment for just the union of the retrieved sets of many different searches for the same topic and assume all unjudged documents are not relevant. The IDEAL collection itself was never constructed, but pooling was used as the methodology to build the first TREC collections.

### 3 TREC Ad Hoc Collections

TREC was conceived as a way of supporting the IR research community by developing the infrastructure necessary to do IR research. It began in 1992 with the goals of creating a single large test collection and standardizing the evaluation measures used to compare test results (Voorhees and Harman 2005). It has accomplished those goals and much more, building scores of test collections for a wide range of search tasks and inspiring other test-collection-building efforts, such as CLEF, that have built yet more collections. Equally as important, the repository of runs collected by the various community evaluations has provided the data needed to empirically examine the soundness of the test collection methodology (Buckley and Voorhees 2005).

### 3.1 Size

The main task in the first 8 years of TREC was the ad hoc task that built eight ad hoc test collections. The ad hoc task is the prototypical Cranfield evaluation task in which systems return a ranked list of documents from a known, static document set for each of a set of previously-unseen topics. What set the TREC ad hoc collections apart from those that had been created earlier was the size of the document set. The collections contain 2–3 gigabytes of text and 500,000–1,000,000 documents. The documents are mostly newswire or newspaper articles, but also include other document types such as government publications to have a heterogeneous mix of subject matter, literary styles and document formats.

As mentioned above, the relevance judgments for these collections were created using pooling. A subset of the runs submitted to TREC in a given year was selected to be the set of judged runs. The number of judged runs was determined such that the total number of documents to be judged fit within the available budget and each TREC participant had an equal number of runs judged (assuming they submitted at least that number). For each judged run, the top  $X$  documents per topic were added to the topics' pools. Most frequently,  $X$  was set to 100. Human assessors then judged each document in the pools where a single assessor judged all the documents for a given topic.

The critical factor in pooling is that unjudged documents are assumed to be not relevant when computing traditional evaluation scores. This treatment is a direct result of the original premise of pooling: that by taking top-ranked documents from sufficiently many, diverse retrieval runs, the pool will contain the vast majority of the relevant documents in the document set. If this is true, then the resulting relevance judgment sets will be “essentially complete”, and the evaluation scores computed using the judgments will be very close to the scores that would have been computed had complete judgments been available.

Various studies have examined the validity of pooling's premise in practice. Harman (1996) and Zobel (1998) independently showed that early TREC collections in fact had unjudged documents that would have been judged relevant had they been in the pools. But, importantly, the distribution of those “missing” relevant documents was highly skewed by topic (a topic that had lots of known relevant documents had more missing relevant), and roughly uniform across runs. Zobel demonstrated that these “approximately complete” judgments produced by pooling were sufficient to fairly compare retrieval runs. Using the leave-out-uniques test, he evaluated each run that contributed to the pools using both the official set of relevant documents published for that collection and the set of relevant documents produced by removing the relevant documents uniquely retrieved by the run being evaluated. The results showed that mean evaluation scores for the runs were only marginally impacted.

Of course, pooling still requires relevance judgments for the documents in the pools, and the reliance on subjective human relevance judgments was the major criticism of the Cranfield methodology from its beginning (Taube 1965; Cuadra and

Katter 1967; Harter 1996). In response to the criticism, Cleverdon examined the effect of using different assessors' judgments when comparing nineteen indexing methods using four independent sets of judgments. When the indexing methods were ranked by score, he found a few differences in the ranks of some methods when varying the judgment sets, but the correlation between the methods' rankings was always very high and the absolute difference in performance of the indexing methods was quite small (Cleverdon 1970). Both Lesk and Salton (1969) and Burgin (1992) also examined the effect of varying judgments on different indexing methods using different collections and found no differences in the relative performance of their respective methods. However, each of these studies was performed on small collections (fewer than 1300 documents) so topics had a correspondingly small number of relevant documents and absolute scores had limited possibility to change. To ensure that stability of retrieval results held for collections with much larger relevant sets, similar tests were repeated on two TREC collections (Voorhees 2000). Those tests included a variety of different conditions including judgments made by query authors vs. judgments by non-authors; judgments made by different non-authors; judgments made by a single judge vs. group judgments; and judgments made by different people in the same environment vs. judgments made in very different environments. In each of these conditions the absolute value of the effectiveness measure was affected by different judgments, but the relative performance of the retrieval runs was almost always the same.

A major result of these ad hoc experiments was the demonstration that the size of a test collection's document set does in fact matter. IR is challenging because of the large number of different ways the same concept can be expressed in natural language, and larger collections are generally more diverse. Further, small collections can have at most a small number of relevant documents for a topic while larger collections can have a much more variable number of relevant documents across topics. Both retrieval system effectiveness and the ability to evaluate retrieval system effectiveness are more challenging when there is greater variability.

### **3.2 *Evaluation Measures***

The second goal for the initial TREC was to standardize evaluation practice, particularly the evaluation measures used to report results. The measures used to evaluate the runs in TREC-1 were measures in common usage right before TREC began. These included precision and recall at various document cut-off levels, interpolated precision at recall points from 0.0 to 1.0 in steps of 0.1 (used to plot a recall-precision graph), and the 11-point average or 3-point average of interpolated precision (averaged over the same recall levels as the recall-precision graph, or at {0.25, 0.5, 0.75} recall, respectively). Two problems became immediately obvious, both related to the number of relevant documents per topic in large collections. First, because different topics had very different numbers of relevant documents, measures based on a constant document cut-off level averaged very poorly. Precision at 30

documents retrieved represents very different retrieval performance for a topic that has only 5 relevant documents as compared to a topic that has 878 (as one of the TREC-1 topics did). Whatever cut-off level is chosen will be appropriate for some topics but wildly inappropriate for others. The second problem was caused by the fact that TREC-1 runs were evaluated over only the top 200 ranks. When topics have more relevant documents than the size of the evaluated set, high recall levels are not obtainable, causing all but the smallest recall level values to be unstable. As an example, consider a topic with 499 relevant documents. A system that retrieves 149 relevant documents in the top 200 ranks never reaches 0.3 recall, so interpolated precision at all recall levels greater than 0.2 is 0.0 (using the interpolation method of trec-eval). However, if the system retrieves 150 relevant documents in the top 200 ranks, its interpolated precision score at 0.3 recall is 0.75 instead of 0.0.

Several changes to the evaluation were therefore instituted for TREC-2 that remained for the rest of the ad hoc tasks. An easy change was to increase the number of documents submitted for a run for a topic from 200 to 1000. Topics were also “narrowed” somewhat such that the target number of relevant documents would be no more than about 350. The idea was that the evaluated set should be at least three times as large as the number of relevant documents to avoid erratic behavior when measuring high recall levels. Another change was to introduce two new evaluation measures, R-precision and noninterpolated average precision. R-precision for a topic with  $R$  relevant documents is precision at rank  $R$ . Noninterpolated average precision, now generally just called average precision, is the mean of the precision at each relevant document over all relevant documents, using 0.0 as the precision of a relevant document not retrieved. When this value is averaged over all topics in a topic set, the result is known as *Mean Average Precision (MAP)*, and it became the single measure most often used in IR research to represent the overall effectiveness of a run.

### 3.3 Reliability Tests

Empirical investigation of the reliability of test collection experiments with respect to two aspects of the methodology—the effect of differences in opinions of relevance and the effect of using essentially complete rather than truly complete judgments sets—was summarized in Sect. 3.1. Here, reliability means that a researcher can have confidence that if an experiment shows that system A is better than system B, then system A will be better than system B in other equivalent environments with high probability. The investigations demonstrated that absolute scores of effectiveness measures change as conditions change, but relative scores are highly consistent. These results underscore an important property of the Cranfield methodology, namely that the only valid use of evaluation scores computed on a test collection is to compare them to other scores computed on the exact same collection. This means, for example, that scores computed on CLEF collections

from two different years are *not* directly comparable, nor are scores computed on a collection and a subset of it.

The collection of runs submitted to various evaluation tasks enabled other empirical studies that help validate the reliability of the Cranfield methodology. One such study examined the size of the collection with respect to the number of topics it contains (Voorhees and Buckley 2002). Another examined the stability of different evaluation measures (Buckley and Voorhees 2000; Sakai 2006). The results of these studies are summarized here.

### 3.3.1 Effect of Topic Set Size

Retrieval system effectiveness has been reported as an average over topics since the first Cranfield experiments because retrieval system performance is known to vary widely depending on the topic. An analysis of variance model fitted to the TREC-3 results demonstrated that the topic and system effects, as well as the interaction between topic and system, were all highly significant, with the topic effect the largest (Banks et al. 1999). What this means is that retrieval effectiveness depends on both which question is asked and which retrieval mechanism is used, but on average which question is asked has a bigger effect on effectiveness than the retrieval mechanism used. Further, different mechanisms work relatively better on different question types.

The set of topics in a test collection is assumed to be a random sample of the universe of possible questions, so there is always some chance that a comparison of two systems using any given test set will lead to the wrong conclusion. The probability of an error can be made arbitrarily small by using arbitrarily many topics, but there are practical limits to the number of topics that can be included in a test collection. While experienced researchers knew that a sufficient number of topics was needed so average scores would be stable, there was little concrete evidence to suggest what was sufficient. The design study for the IDEAL collection posited that fewer than 75 topics would not be useful (Spärck Jones and Van Rijsbergen 1975). TREC organizers, who had to balance cost of topic development and relevance judgments against the quality of the collection, chose 50 topics as the default size for the TREC ad hoc collections.

Voorhees and Buckley (2002) used TREC results to empirically derive collection error rates. An error rate is defined as the likelihood of reaching a wrong conclusion from a single comparison as a function of the number of topics used in the comparison and the size of the difference of the evaluation scores (called  $\Delta$ ). Once established, the error rates were used to derive the minimum difference in scores required for a certain level of confidence in the results given the number of topics used in the comparison.

The core of the procedure used to estimate the error rates was comparing the effectiveness of a pair of runs on two disjoint topic sets of equal size to see if the two sets disagreed as to which of the runs is better. The comparisons were repeated for many different pairs of runs and many different topic sets. The error rate is defined

as the percentage of times that the two topic sets disagreed as to which is the better system. Since TREC runs contain 50 topics, this procedure was used to directly compute error rates for topic set sizes up to 25. Curves of the form  $\text{ErrorRate} = A_1 e^{-A_2 S}$  where  $S$  is the size of the topic set were fit to the observed error rates, and then those curves were used to extrapolate error rates for larger topic sets. A different curve was fit for each of a set of binned  $\Delta$  values. As expected, error rates are larger for smaller  $\Delta$ 's and decrease as the number of topics increases.

Spärck Jones (1974) suggested the rule-of-thumb that differences in scores of 0.05 were noticeable and differences of 0.1 were material (for small collections and using measures other than MAP). For MAP and topic set sizes of 25, the error rate computed over the TREC collections for a difference of 0.05 is approximately 13% on the TREC ad hoc collections. This means that if we knew nothing about systems A and B except their MAP scores which differed by 0.05, and if we repeated the experiment on 100 different sets of 25 topics, then on average we would expect 13 out of those 100 sets to favor one system while the remaining 87 would favor the other. The error rate for a difference of 0.1 with 25 topics is much smaller at approximately 2.5%. The error rates are also much smaller for sets of 50 topics, 3.7% and 0.15% respectively. For topic sets of 50 topics, a difference of 0.05 was the smallest  $\Delta$  with an error rate less than 5%.

These differences in MAP scores used to compute the error rates are *absolute* differences, while much of the IR literature reports *percentage* differences. An absolute difference of 0.1 is a very substantial difference, especially given that the best MAP scores on the TREC ad hoc collections are approximately 0.3. The percentage difference between a run with a 0.3 MAP score and a run with a 0.10 absolute difference is approximately 33% and for a 0.05 absolute difference is approximately 15%. However, the computed error rates are also for a single comparison of two arbitrary runs. In practice, researchers will use multiple test collections to compare different techniques, and the techniques being compared will likely be variants of some common system. Comparisons of different instances of a common system will have less variability overall, so error rates will be smaller in this case. Using multiple test collections is sound experimental practice and will again increase the confidence in conclusions reached.

### 3.3.2 Effect of Evaluation Measure Used

The study of the effect of the topic set size summarized above showed that the reliability of experimental findings depends on (at least) three interrelated components of the Cranfield paradigm: the number of topics used, the evaluation measure used, and the difference in scores required to consider one method better than the other. The evaluation measure used makes a difference because measures have different inherent reliabilities. Buckley and Voorhees (2000) focused on quantifying these differences among measures using the TREC Query Track (Buckley 2001) data.

The Query Track data provides different expressions of the same underlying information need. That is, in TREC parlance, the track gathered different queries for

the same topic and ran different retrieval systems on each of the different queries. Each query provides a separate evaluation score for the corresponding topic, thus producing a set of scores for the exact same topic. While using different queries does affect retrieval behavior—some queries are clearly better expressions of the topic than others—the effect of the number of relevant documents on system behavior is controlled because it remains constant. Controlling this topic effect allows the error inherent in the evaluation measure itself to be isolated.

Call a *query set* a collection of 50 queries, one for each topic. Each of 21 query sets was run using nine different retrieval methods, producing a data set consisting of nine sets of the top 1000 documents retrieved for each of 1050 queries (21 versions of 50 topics).

As in the topic set size experiment, error rates for an evaluation measure are computed by comparing the scores obtained by different retrieval methods, but the particulars of how the error rate is defined differ. Buckley and Voorhees (2000) used the error rate calculation described here, while Sakai (2006) used a separate, more mathematically-principled definition, with both definitions leading to the same conclusions. The first approach counts the number of times each retrieval method was better than, worse than, and equal to each other retrieval method when compared over a given query sets, using many different permutations of queries assigned to query sets and considering scores within a given percentage difference (say, 5%) of one another to be equivalent. Assuming that the correct answer is given by the greater of the better-than and worse-than values, the lesser of those two values is the number of times a test result is in error. Hence the error rate is defined as the total number of errors across all method pairs divided by the total number of decisions. With this definition, the error rate can never be more than 50%, and random effects start dominating the calculation of the error rate if it exceeds approximately 25%. The number of times methods are deemed to be equivalent is also of interest because it reflects on the power of a measure to discriminate among systems. It is possible for a measure to have a low error rate simply because it rarely concludes that two methods are different. The proportion of ties, defined as the total number of equal-to counts across all method pairs divided by the total number of decisions, quantifies this effect.

The error rates for different measures were found to be markedly different. Measures that depend on a relatively few highly ranked documents, such as precision at small cut-off levels, have higher error rates than measures that incorporate more documents. For example, when using a fuzziness factor of 5%, Prec(10) and Prec(30) had error rates of 3.6% and 2.9% respectively, while MAP had an error rate of 1.5%. The proportion of ties for the various measures also differed substantially. Precision failed to distinguish between two systems almost a quarter of the time (24%) while MAP failed to distinguish about 13% of the time.



### 3.3.3 Significance Testing

The error rates computed in the two investigations described earlier in this section are different from statistical significance tests, but all acknowledge the same underlying truth of test-collection-based experiments: that there is a fair amount of noise in the process. Statistical significance tests are run on the results of a retrieval experiment to determine whether the observed variation in topic scores is consistent with chance fluctuations.

Statistical significance testing has been used in IR experiments for almost as long as test collections have existed (Lesk 1967), though their application in retrieval experiments has not been without controversy. Early critics were concerned that retrieval system output does not meet the distributional assumptions of parametric tests (Van Rijsbergen 1979). Proponents demonstrated that the test were robust to the types of violations seen in practice (Hull 1993; Smucker et al. 2007) or suggested non-parametric schemes such as the bootstrap method (Savoy 1997). More recent concerns have arisen because of the wide availability of test collections, especially collections with very many topics. The wide availability of test collections means that it is easy to run experiments: a wide variety of different techniques can all be compared to one another, but corrections for multiple comparisons are seldom used (Carterette 2012). Further, given that the field (re)uses the same collections, there are also sequential testing effects (Carterette 2015). Sakai (2016) provides a survey of current practices in significance testing in IR.

The final test of the validity of the Cranfield paradigm is whether the conclusions reached from the laboratory experiments transfer to operational settings. Hersh and his colleagues suggest that the results may not transfer since they were unable to verify the conclusions from a laboratory experiment in either of two user studies (Hersh et al. 2000; Turpin and Hersh 2001). However, their tests were small and the user studies did not show that the conclusions from the laboratory test were wrong, simply that the user studies could not detect any differences. Furthermore, using a different approach Al-Maskari et al. (2008) demonstrated that users were indeed able to discern and act on the differences found in systems whose test-collection-based scores were only slightly different. Even a cursory examination of retrieval technology actually in use today makes it clear that the results do transfer. Basic components of current web search engines and other commercial retrieval systems—including full text indexing, term weighting, and relevance feedback—were first developed on test collections.

## 4 Moving On

TREC was founded on the belief that the Cranfield paradigm of using test collections as laboratory tools to compare the effectiveness of different retrieval methods was fundamentally sound though in need of updating with regard to collection size and standardization of evaluation metrics. Research using subsequently constructed

collections and retrieval results confirmed this belief, as summarized above. Yet those findings apply to a fairly narrowly proscribed protocol that is not strictly applicable to much of IR research in the ensuing years.

This section looks at ways in which the Cranfield paradigm has been extended or modified to continue to support the IR research community. In keeping with the scope of the chapter, the section only focuses on evaluation protocols connected to some form of a test collection. Protocols for controlled experiments involving users of operational systems, including traditional interactive IR studies and newer online evaluation (e.g., A/B testing and reuse of data gleaned from query logs) experiments, have also evolved in the ensuing years, but are not discussed here. Kelly (2009) provides a comprehensive review of interactive IR and Hofmann et al. (2016) provides the same for online evaluation.

## ***4.1 Cross-Language Test Collections***

A cross-language ad hoc retrieval task was the inaugural task in CLEF and is also featured prominently in NTCIR and FIRE. As an ad hoc retrieval task, Cranfield is clearly an appropriate evaluation tool for it, but building a good cross-language test collection is much more difficult than building a monolingual collection.

When creating a cross-language collection, a topic will be created in an initial language, and then usually translated into some of the other languages of the document set (to facilitate multiple monolingual experiments or cross-language experiments with differing source languages, for example). The quality of this translation is very important: a too literal translation depresses retrieval results because the language use in the translated topic does not match how the concept is natively expressed in the documents (Mandl and Womser-Hacker 2003).

Even with good translations, a given topic is much more likely to pertain to some parts of the collection than others since cultural differences make some topics more apt to be discussed in some subset of languages. This complicates pooling for cross-language collections. The quality of a test collection depends on having diverse pools, yet it is very difficult to get equally large, diverse pools for all languages contained within a multilingual collection. Both the number of runs submitted by participants and the documents retrieved within a run are usually skewed in favor of some languages at the expense of others. As a result, the pools for the minority languages are smaller and less diverse than the pools for the majority languages, which introduces an unknown bias into the judgments. Ensuring an equal number of documents is judged per language is not a solution to this problem because of the inherent differences in the true number of relevant documents per language. One way that does help enhance the quality of the pools is for the collection builders to supplement pools built from participant runs with documents discovered through the builders' own manual searches, a technique used to good advantage for the early NTCIR collections (Kando et al. 1999).

Obtaining a consistent set of relevance judgments is also more difficult for cross-language collections. In monolingual collections, the judgments for a topic are produced by one assessor. While this assessor's judgments may differ from another assessor's judgments, the judgment set represents an internally consistent sample of judgments. Using a single individual to judge documents across the multiple different languages represented in a cross-language collection is generally infeasible, however. Instead, cross-language collections are typically produced using a separate set of assessors for each language, and thus multiple assessors judge the same topic across the entire collection. This necessitates close coordination among assessors so that different cultural understandings of the topic can be resolved and the typical "gray areas" of relevance can be judged consistently across languages.

## ***4.2 Other Tasks***

The document ranking abstraction that is the basis of the standard Cranfield paradigm is applicable to many information access tasks. The abstraction (though not the technical solutions) is independent of document type, including not only various textual genres but other media types such as recordings of speech or images or videos. The abstraction is also independent of the expression of the information need, such as using a natural language statement, a structured query, or a sample relevant document. The abstraction applies whenever the actual user task involves a searcher interacting with a set of distinct, uniquely identified information units (the documents) returned in response to the searcher's request. Nonetheless, there are a number of realistic information access tasks that do not fit this precise abstraction. This section describes how standard Cranfield has been modified to support research for three other families of abstract tasks: filtering, focused retrieval, and web-based search.

### **4.2.1 Filtering Tasks**

If ad hoc searching is thought of as "pull" technology where the user pulls documents from the system by querying, filtering is "push" technology where the system periodically informs the user of a new document. In the abstract filtering task, the topics of interest are relatively stable and are known in advance; the system task is to find relevant documents for each topic from a document stream (such as a newswire or social media feed). The main distinguishing feature of a filtering task is that the system must make a binary decision for each document in the stream as to whether that document will be returned to the user for the current topic, and that decision must be made relatively shortly after the document appears in the stream. Making a binary decision is a strictly more difficult task than ranking (Robertson and Callan 2005).

In a typical filtering task, systems receive feedback in the form of a relevant judgment for documents they retrieve, and adapt their processing based on the judgments. This makes set-based precision and recall measures inappropriate to evaluate system performance just as ranked-retrieval evaluation measures are clearly inappropriate. Filtering tasks are generally evaluated using some sort of utility measure, where systems are rewarded a gain for retrieving a relevant document and penalized a loss for retrieving a non-relevant document. Latency, the amount of time between when the first document appears in the stream and the decision to retrieve it can also be incorporated into the measure (Aslam et al. 2014).

Building test collections for filtering tasks requires having a document set that has a well-defined order to the document stream; generally such an order is related to time. To support adaptive filtering, the relevant set must be known prior to system execution, meaning traditional pooling based on participant runs is not an option. The TREC 2002 Filtering track compared two ways of building such a collection: using several rounds of relevance feedback searches during topic development, and using category descriptors from the document source as a kind of judgment (Soboroff and Robertson 2003). The results from the track demonstrated that these two types of collections were quite different in how they ranked systems.

#### 4.2.2 Focused Retrieval Tasks

Focused retrieval is a general category of tasks in which documents are no longer treated as atomic entities (Trotman et al. 2010). This broad category of tasks includes passage retrieval such as in the TREC HARD track (Allan 2003); question answering such as in the TREC QA track (Voorhees 2005); and XML element retrieval as studied in *INitiative for the Evaluation of XML Retrieval (INEX)* (Bellot et al. 2014). The task abstraction for focused retrieval tasks generally maintains the ad hoc nature and ranking aspects of Cranfield, but systems do not retrieve distinct, uniquely identified information units. This means evaluation schemes can no longer use simple matching between a gold standard answer (e.g., judgment files) and system results.

Passage retrieval evaluation generally requires matching system-returned document extracts to a set of standard relevant passages. Since it is unlikely that systems will return extracts that match exactly, strict pooling to find a set of gold-standard relevant passages is not possible. Further, since systems are also unlikely to return extracts that match the gold standard passages exactly, evaluation measures must account for redundancy and omissions in the returned passages. The TREC HARD used a function of character-level recall and precision with respect to a set of gold-standard relevant passage extracts.

Question answering system evaluation shares the same problem as passage retrieval in that the answers returned by different systems are seldom exactly the same. Further, automatically determining whether a system response contains a correct answer is generally as difficult as the question answering task itself. To create a form of reusable test collection for the short-answer, factoid question in the initial

TREC QA track, organizers (manually) created regular expression patterns from the set of pooled system responses, and treated a new answer string as correct if and only if the string matched a pattern.

XML elements have unique ids (the path from the document root to the element is a unique id), but granularity is an issue for both defining the set of gold-standard relevant elements as well as matching system output to the relevant elements set. Intuitively, a system should receive some credit for retrieving an element that contains a relevant element, but only if the containing element is not too large (e.g., no credit for retrieving an entire book if the relevant element is one small element in a single sub-sub-section of a single chapter). Similarly, a system should receive credit if it retrieves a too narrow element, assuming the narrow element is comprehensible on its own. Balancing such considerations while also accounting for redundancy (e.g., not giving double credit for retrieving two elements each of which contains the same relevant sub-element) to accurately model when one system response is better than another is quite challenging. INEX has judged relevance on two scales, exhaustivity and specificity, and combined those judgments using a form of cumulated gain; see Lalmas and Tombros (2007) for details.

### 4.2.3 Web Tasks

While web search can be construed as an ad hoc search task over documents that happen to be web pages, Cranfield is not a good abstraction of web search for several reasons. Cranfield is an abstraction of informational search as befits its library heritage while people use the web in other ways including using search for navigation and for transactions (Broder 2002). Further, it is not clear what “the document set” is for a web test collection. Consider a web page that contains code that dynamically generates the content seen by a visitor to it. Is the document the static code? the entire data environment that determines the content seen at any given time? the particular content presented to some one visitor? The lack of editorial control gives rise to spam documents making the web the rare corpus in which the words of the documents themselves are *not* necessarily a good indicator of document content. The enormity of the web and its transitory nature precludes a classic static test collection that meaningfully represents general web search.

Particular aspects of the overall web search problem have been studied using test collections, however (Hawking and Craswell 2005). These efforts have been supported by specific crawls that gathered a coherent subset of the web at a given point of time<sup>1</sup> and which were then used with queries drawn from contemporaneous internet search engine logs.

Early editions of the TREC Web Track studied home page and named page finding, navigational search tasks. One outcome of this work was demonstrating

---

<sup>1</sup>See [http://ir.dcs.gla.ac.uk/test\\_collections/access\\_to\\_data.html](http://ir.dcs.gla.ac.uk/test_collections/access_to_data.html) and <https://lemurproject.org/clueweb12/>.

the effectiveness of using anchor text to support navigational search (Hawking and Craswell 2005). Navigational search is generally evaluated using either reciprocal rank (the reciprocal of the rank at which the correct URL was retrieved) or “success at  $n$ ”, a binary measure that signifies whether the correct URL was retrieved in the top  $n$  ranks.

Given the size of the web and the brevity of the typical web search query, there is often a spectrum of information needs that the query might represent. Sometimes the query is inherently ambiguous; other times it may refer to a single broad area of interest that has multiple distinct aspects. For example, the query *avp* is ambiguous in that it might refer to the Wilkes-Barre Scranton International Airport (airport code AVP), the Avon Products Company (stock symbol AVP), the “Alien vs. Predator” movie franchise, AVP antivirus software, or the Association of Volleyball Professionals. The query *moths* likely refers to the winged insects, but the actual information need could be a desire to see pictures of moths in general; identification of a specific instance of a moth; controlling a moth infestation; distinguishing between moths and butterflies; understanding moth habitats; etc. Web pages that are excellent documents for one aspect may be completely irrelevant for another.

Diversity web tasks look to develop search systems that are able to cover the different aspects of a query within the top results. The evaluation methodology that supports diversity tasks requires a delineation of the aspects to be covered by a query statement, relevance judgments for each such aspect for each judged page, and evaluation measures that appropriately reward ranked lists for coverage as well as accuracy. Two such measures are  $\alpha$ -NDCG (Clarke et al. 2008) and ERR-IA (Chapelle et al. 2011). Clarke et al. (2011) show that these measures behave as intended, rewarding systems that achieve a balance between novelty and overall precision.

### 4.3 *Size Revisited*

While TREC ad hoc collections contained much bigger document sets than the collections generally available at the start of TREC, the ad hoc collections are once again quite small compared to many document sets that are searched by operational systems—including, but not limited to, the web. Unfortunately, pooling has its own size dependency that prevents its successful application for arbitrarily large document sets. Pooling’s fundamental assumption that the pools contain an unbiased sample of the relevant documents becomes untenable unless the size of the pools grows in concert with the size of the document set. Otherwise, the sheer number of documents of a certain type (for example, relevant documents that contain many query words) fill up the pools to the exclusion of other types of documents (for example, relevant documents that contain few query words) (Buckley et al. 2007). Systems that are able to retrieve the minority type of relevant documents are unfairly penalized when evaluated by the relevance judgments produced by shallow pools.

One way of adapting Cranfield to accommodate these larger document collection sizes is to explicitly acknowledge unjudged documents and account for them in the evaluation. Most frequently, this accommodation has been through the use of evaluation measures specifically designed for partial judgment sets.

### 4.3.1 Special Measures

Buckley and Voorhees (2004) introduced the *bpref* (“binary preference”) measure as a means of evaluating retrieval systems when the relevance judgment sets are known to be far from complete. Sakai and Kando (2008) investigated the fidelity of evaluation results for incomplete judgments by comparing *bpref* and standard evaluation measures computed over only judged documents. That is, they computed evaluation scores by removing unjudged documents from the ranking rather than assuming those documents were not relevant and called these compressed-list versions *measure'*. Among other findings, they concluded that *bpref* was inferior to *MAP'* in terms of both defining the set of statistically different run pairs and the overall similarity of runs ranked by effectiveness as measured by Kendall's  $\tau$  correlation. To test the measures, they produced increasingly incomplete judgment sets by taking random subsets of the original judgment sets for existing test collections. Sakai subsequently showed that realistic judgment set building is subject to both system and pool depth biases (Sakai 2008a,b) that are not modeled well by random subsets. In more realistic scenarios, the compressed list versions of the measures had no clear advantage over the traditional versions, though they were superior to *bpref*.

*Bpref* and compressed list versions of standard measures can be computed using any existing test collection. A family of measures known as *inferred measures* are available if the test collection is constructed to support them. *Inferred measures* are defined as statistical estimates of the true value of the corresponding traditional measures (Yilmaz and Aslam 2008).

As an example, *inferred AP* computes an estimate of the expectation of the following random experiment when assuming the known (incomplete) set of relevance judgments is a uniform random sample of the complete judgment set. Given a retrieval result of a ranked list for a single topic:

1. Select a relevant document at random from the collection. Call the rank of this relevant document in the retrieved list  $k$ .
2. Select a rank  $i$  at random from among the set  $\{1, \dots, k\}$ .
3. Output the binary relevance of the document at rank  $i$ .

Under the assumption that a uniform random sample of the relevant documents is known, mean *inferred AP* is a good estimate of the actual value of *MAP*. However, in practice, incomplete judgment sets are seldom uniform random samples of the complete set—relevant documents retrieved higher in ranked lists are more likely to be included in the known set, for example. *Inferred measures* were thus extended to a collection-building method that samples from the runs in such a way as to

maintain accurate estimates of the measures' values, producing extended inferred measures (Yilmaz et al. 2008). The extended inferred measures technique builds judgment sets using stratified random sampling across the run set. That is, judgment sets to be used in computing extended inferred measures are created by taking uniform random samples of different regions of the ranked document lists where the different regions are sampled at different rates. The particular sampling strategy used affects the quality of the resulting estimates. Effective strategies do not include large, sparsely-sampled strata and do include a small top stratum that is exhaustively judged (e.g. depth-10 pools) (Voorhees 2014).

Much of the difficulty of getting fair evaluation results for large collections lies in getting good estimates of the number of relevant documents,  $R$ . The TREC Legal Track, which focused on the problem of legal discovery where *relevant* document sets can be very large, used stratified sampling (which differed from inferred measure sampling) to estimate  $R$  (Tomlinson and Hedin 2011). In this case, the strata were defined using the number of runs that retrieved a document. Others contend that  $R$  (and thus recall) is not a good basis for retrieval system evaluation since it is unimportant (relevant set sizes are too large to be meaningfully processed by a user) and unknowable. Moffat and Zobel (2008) introduced *Rank-Biased Precision (RBP)* as a measure that does not rely on knowledge of  $R$  and whose true value can be bounded in the presence of incomplete judgments.

RBP is based on a user model that assumes a user starts at the top of a ranked list and proceeds to the next document with probability  $p$ . The measure is defined as the expected rate at which relevant documents are found conditioned on  $p$ :

$$\text{RBP} = \sum_{i=1}^n \text{rel}_i (1 - p)^{i-1} p$$

where  $n$  is the number of ranks over which the score is computed. The RBP score for a ranking that is a prefix of another ranking is by definition a lower bound for the RBP score of the extended ranking. This property means that upper and lower bounds for RBP for a given ranking in the presence of unjudged documents are easy to compute: the lower bound is the RBP score when all unjudged documents are treated as not relevant, and the upper bound is the score when all unjudged documents are treated as relevant. Larger differences between the two bounds, caused by encountering more unjudged documents, are an indicator of greater uncertainty in the evaluation.

### 4.3.2 Constructing Large Collections

The stratified sampling used to support different evaluation measures is one example of modified collection construction techniques in support of building fair larger collections. Other construction techniques not tied to particular measures have also been tried.



Since the number of human judgments that can be obtained is usually the limiting factor on the size of a test collection, several approaches look at different ways of allocating assessor resources. Cormack et al. (1998) introduced the *Move-to-Front (MTF)* method of selecting the next document to judge from a set of runs based on the relevance judgments already received. The method favors selecting additional documents from runs that have recently retrieved relatively many relevant documents. Losada et al. (2016) showed that MTF is one instance of a family of multi-arm bandit document selection techniques. Each bandit method tries to maximize the number of relevant documents found while staying within a given budget of judgments, but differ in the details of precisely how the next run to contribute a document is selected. The most effective bandit methods are dynamic methods like MTF that require relevance judgments on previous selections before selecting the next document. The exploration of MTF and other bandit methods has used simulation on existing collections to show that the vast majority of relevant documents can be recovered with many fewer judgments than pooling required. Implementing dynamic methods to build a new collection from scratch is logistically more difficult than pooling, and also potentially adds assessor bias to the assessments since assessors know they are seeing documents in quality order.

Another choice when allocating assessor resources is balancing the number of topics in the test set against the exhaustiveness of the judgments for those topics. Conventional wisdom such as in the IDEAL collection report is that more topics are always better, but that is assuming essentially complete judgments for each topic. Sanderson and Zobel (2005) found that many shallowly-judged topics (whose runs were thus evaluated using precision-based measures) resulted in collections that ranked systems more similarly to an existing high-quality collection than fewer topics with deeper judgments. However, several studies have also shown that you can find small subsets of topics from a larger collection that rank systems the same as the full collection (Guiver et al. 2009; Hosseini et al. 2012). Kutlu et al. (2018) reconciles these differing findings by showing that many shallowly-judged topics are better when topics are chosen at random, but for smaller budgets and when topic development costs are comparatively high, using a selectively chosen, smaller set of more thoroughly judged topics produces a more reliable collection that is also more reusable.

The *Minimal Test Collection (MTC)* protocol is a dynamic method that can be used to create a collection being built to compare a specific set of runs known at collection build time (Carterette et al. 2006). MTC identifies those documents that will best distinguish the set of runs under some measure. Empirically, using MAP as the focus measure creates a collection that also fairly compares the run set using other measures such as  $Prec(10)$  or R-precision.

#### 4.4 User-Based Measures

Research on evaluation measures used with test collections has not focused solely on accommodating incomplete relevance judgments. Another focus of measure work has been incorporating different models of search behavior into the measures. This section summarizes some of work in this area; also see Sakai (2014) for an alternative summary.

RBP was introduced above as a measure that accommodates partial judgments, but it also codifies a specific user model of a searcher traversing a ranked list from the top who proceeds to the next rank in the list with probability  $p$  (Moffat and Zobel 2008). ERR-IA and  $\alpha$ -NDCG, also mentioned earlier, similarly assume what Clarke et al. (2011) call a cascade model of user behavior: considering the relationship between successive elements of a result list. Indeed,  $\alpha$ -NDCG is an extension of the *Discounted Cumulated Gain (DCG)* family of measures that explicitly encodes the cascade user model and also accommodates different grades of relevance (Järvelin and Kekäläinen 2002).

The user model for DCG is again a searcher traversing a ranked list from the top, this time proceeding to a fixed rank  $k$ . At each rank, the searcher accumulates a gain proportional to the relevance grade of the document at the rank, with the base amount of gain for a given relevance grade reduced in proportion to the depth of the rank. While there are several different formulations of the measure, a frequently used definition (Burges et al. 2005) is

$$\text{DCG}_k = \sum_{i=1}^k \frac{2^{\text{rel}_i} - 1}{\log_2(i + 1)}$$

where  $k$  is the rank to which the searcher traverses and  $\text{rel}_i$  is the gain value for the relevance grade of the document at rank  $i$ . The rank-related penalty arises from the logarithm in the denominator. The base of the logarithm is a parameter of the measure and models the persistence of the searcher reviewing the ranked list.

The maximum DCG score for a topic depends on the number of relevant documents of each grade and the gains associated with the grades. This means the maximum score across topics varies widely and hence DCG does not average well, so must be normalized producing *normalized Discounted Cumulative Gain (nDCG)*. Each topic's DCG score is normalized by dividing it by the maximum score possible for the topic. The maximum is obtained by scoring a ranked list that contains all of the documents with maximum relevant grade first, followed by all relevants with the next highest relevance grade, and so forth until all relevant documents are ranked.

Carterette (2011) shows that both nDCG and RBP, as well as more traditional measures such as average precision and reciprocal rank, can all be modeled using

sums over the product of a discount function of ranks and a gain function that maps relevance judgments to numeric utility values:

$$M = \sum_{k=1}^K \text{gain}(\text{rel}_k) \times \text{discount}(k).$$

He then shows that any such measure is actually a composition of three independent component models, the model that describes how the user interacts with the results called the browsing model; the model that describes the utility the user obtains from an individual document called the document utility model; and the model that describes how utility is accumulated while browsing, the utility accumulation model. By relating measures that share a common component model, the framework can unify previously disparate measures into a small set of measure families, as well as suggest new measures that would fill previously unoccupied areas of the measure space. One outcome of the initial investigation into the measure space was a demonstration that DCG is a robust measure that does in fact model user-centered behavior.

## 5 Conclusion

The Cranfield paradigm has proved to be remarkably resilient. Despite fierce criticism from the start and periodic pronouncements of its impending demise, the paradigm has enabled research that has greatly improved retrieval performance in practice. This success has largely resulted *because* of the paradigm's limitations rather than despite them. The document ranking task is a carefully calibrated level of abstraction that has sufficient fidelity to real user tasks to be informative, but is sufficiently abstract to be broadly applicable, feasible to implement, and comparatively inexpensive. By eliminating anything that does not directly contribute to the core competency, Cranfield loses realism but gains substantial experimental power.

Maintaining a proper tension between realism and abstraction is key to extending the paradigm to new tasks. It obviously does no good to abstract an evaluation task to the point where test results do not reflect performance on the real task of interest; it is equally as unhelpful to include any operational variable that might possibly influence outcomes since generalization then becomes impossible and nothing is learned.

## References

- Al-Maskari A, Sanderson M, Clough P, Airio E (2008) The good and the bad system: does the test collection predict users' effectiveness? In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '08. ACM, New York, pp 59–66
- Allan J (2003) HARD Track overview in TREC 2003: high accuracy retrieval from documents. In: Proceedings of the twelfth Text REtrieval Conference (TREC 2003)
- Aslam J, Ekstrand-Abueg M, Pavlu V, Diaz F, McCreddie R, Sakai T (2014) TREC 2014 temporal summarization track overview. In: Proceedings of the twenty-third Text REtrieval Conference (TREC 2014)
- Banks D, Over P, Zhang NF (1999) Blind men and elephants: six approaches to TREC data. *Inf Retr* 1:7–34
- Bellot P, Bogers T, Geva S, Hall MA, Huurdeman HC, Kamps J, Kazai G, Koolen M, Moriceau V, Mothe J, Preminger M, SanJuan E, Schenkel R, Skov M, Tannier X, Walsh D (2014) Overview of INEX 2014. In: Kanoulas E, Lupu M, Clough P, Sanderson M, Hall M, Hanbury A, Toms E (eds) Information access evaluation – multilinguality, multimodality, and interaction. Proceedings of the fifth international conference of the CLEF initiative (CLEF 2014). Lecture notes in computer science (LNCS), vol 8685. Springer, Heidelberg, pp 212–228
- Broder A (2002) A taxonomy of web search. *SIGIR Forum* 36(2):3–10
- Buckley C (2001) The TREC-9 query track. In: Voorhees E, Harman D (eds) Proceedings of the ninth Text REtrieval Conference (TREC-9), pp 81–85
- Buckley C, Voorhees EM (2000) Evaluating evaluation measure stability. In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR 2000, pp 33–40
- Buckley C, Voorhees EM (2004) Retrieval evaluation with incomplete information. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, pp 25–32
- Buckley C, Voorhees EM (2005) Retrieval system evaluation. In: Voorhees EM, Harman DK (eds) TREC: experiment and evaluation in information retrieval. MIT Press, Boston, chap 3, pp 53–75
- Buckley C, Dimmick D, Soboroff I, Voorhees E (2007) Bias and the limits of pooling for large collections. *Inf Retr* 10:491–508
- Burges C, Shaked T, Renshaw E, Lazier A, Deeds M, Hamilton N, Hullender G (2005) Learning to rank using gradient descent. In: Proceedings of the 22nd international conference on machine learning, ICML '05. ACM, New York, pp 89–96
- Burgin R (1992) Variations in relevance judgments and the evaluation of retrieval performance. *Inf Process Manag* 28(5):619–627
- Carterette B (2011) System effectiveness, user models, and user utility: a conceptual framework for investigation. In: Proceedings of the 34th International ACM SIGIR conference on research and development in information retrieval (SIGIR'11). ACM, New York, pp 903–912
- Carterette BA (2012) Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Trans Inf Syst* 30(1):4:1–4:34
- Carterette B (2015) The best published result is random: Sequential testing and its effect on reported effectiveness. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, SIGIR '15, pp 747–750
- Carterette B, Allan J, Sitaraman R (2006) Minimal test collection for retrieval evaluation. In: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, pp 268–275
- Chapelle O, Ji S, Liao C, Velipasaoglu E, Lai L, Wu SL (2011) Intent-based diversification of web search results: metrics and algorithms. *Inf Retr* 14(6):572–592

- Clarke CL, Kolla M, Cormack GV, Vechtomova O, Ashkan A, Büttcher S, MacKinnon I (2008) Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '08. ACM, New York, pp 659–666
- Clarke CL, Craswell N, Soboroff I, Ashkan A (2011) A comparative analysis of cascade measures for novelty and diversity. In: Proceedings of the fourth ACM international conference on web search and data mining, WSDM '11. ACM, New York, pp 75–84
- Cleverdon CW (1967) The Cranfield tests on index language devices. In: Aslib proceedings, vol 19, pp 173–192, (Reprinted in *Readings in Information Retrieval*, K. Spärck-Jones and P. Willett, editors, Morgan Kaufmann, 1997)
- Cleverdon CW (1970) The effect of variations in relevance assessments in comparative experimental tests of index languages. Tech. Rep. Cranfield Library Report No. 3, Cranfield Institute of Technology, Cranfield, UK
- Cleverdon CW (1991) The significance of the Cranfield tests on index languages. In: Proceedings of the fourteenth annual international ACM/SIGIR conference on research and development in information retrieval, pp 3–12
- Cormack GV, Palmer CR, Clarke CLA (1998) Efficient construction of large test collections. In: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '98. ACM, New York, pp 282–289
- Cuadra CA, Katter RV (1967) Opening the black box of relevance. *J Doc* 23(4):291–303
- Gilbert H, Spärck Jones K (1979) Statistical bases of relevance assessment for the 'IDEAL' information retrieval test collection. Tech. rep., Computer Laboratory, University of Cambridge. Available at <http://sigir.org/resources/museum/>
- Guiver J, Mizzaro S, Robertson S (2009) A few good topics: experiments in topic set reduction for retrieval evaluation. *ACM Trans Inf Syst* 27(4):21:1–21:26
- Harman D (1996) Overview of the fourth Text REtrieval Conference (TREC-4). In: Harman DK (ed) Proceedings of the fourth Text REtrieval Conference (TREC-4), pp 1–23, nIST Special Publication 500-236
- Harter SP (1996) Variations in relevance assessments and the measurement of retrieval effectiveness. *J Am Soc Inf Sci* 47(1):37–49
- Hawking D, Craswell N (2005) The very large collection and web tracks. In: Voorhees EM, Harman DK (eds) TREC: experiment and evaluation in information retrieval. MIT Press, Boston, chap 9, pp 199–231
- Hersh W, Turpin A, Price S, Chan B, Kraemer D, Sacherek L, Olson D (2000) Do batch and user evaluations give the same results? In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR 2000, pp 17–24
- Hofmann K, Li L, Radlinski F (2016) Online evaluation for information retrieval. *Found Trends Inf Retr* 10(1):1–117
- Hosseini M, Cox IJ, Milic-Frayling N, Shokouhi M, Yilmaz E (2012) An uncertainty-aware query selection model for evaluation of IR systems. In: Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval, SIGIR '12. ACM, New York, pp 901–910
- Hull D (1993) Using statistical testing in the evaluation of retrieval experiments. In: Proceedings of the 16th annual international ACM/SIGIR conference on research and development in information retrieval, SIGIR '93. ACM, New York, pp 329–338
- Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of IR techniques. *ACM Trans Inf Syst* 20(4):422–446
- Kando N, Kuriyama K, Nozue T, Eguchi K, Kato H, Hidaka S (1999) Overview of IR tasks at the first NTCIR workshop. In: Proceedings of the first NTCIR workshop on research in Japanese text retrieval and term recognition, pp 11–44
- Keen EM (1966) Measures and averaging methods used in performance testing of indexing systems. Tech. rep., The College of Aeronautics, Cranfield, England. Available at <http://sigir.org/resources/museum/>

- Kelly D (2009) Methods for evaluating interactive information retrieval systems with users. *Found Trends Inf Retr* 3(1–2):1–224
- Kutlu M, Elsayed T, Lease M (2018) Learning to effectively select topics for information retrieval test collections. *Inf Process Manag* 54(1):37–59
- Lalmas M, Tombros A (2007) Evaluating XML retrieval effectiveness at INEX. *SIGIR Forum* 41(1):40–57
- Ledwith R (1992) On the difficulties of applying the results of information retrieval research to aid in the searching of large scientific databases. *Inf Process Manag* 28(4):451–455
- Lesk ME (1967) SIG – the significance programs for testing the evaluation output. In: *Information storage and retrieval*, Scientific Report No. ISR-12, National Science Foundation, chap II
- Lesk M, Salton G (1969) Relevance assessments and retrieval system evaluation. *Inf Storage Retr* 4:343–359
- Losada DE, Parapar J, Barreiro A (2016) Feeling lucky?: multi-armed bandits for ordering judgements in pooling-based evaluation. In: *Proceedings of the 31st annual ACM symposium on applied computing, SAC '16*. ACM, New York, pp 1027–1034
- Mandl T, Womser-Hacker C (2003) Linguistic and statistical analysis of the clef topics. In: Peters C, Braschler M, Gonzalo J, Kluck M (eds) *Advances in cross-language information retrieval: third workshop of the cross-language evaluation forum (CLEF 2002) revised papers*. Lecture notes in computer science (LNCS), vol 2785. Springer, Heidelberg, pp 505–511
- Moffat A, Zobel J (2008) Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans Inf Syst* 27(1):Article 2
- Robertson S (2008) On the history of evaluation in IR. *J Inf Sci* 34(4):439–456
- Robertson S, Callan J (2005) Routing and filtering. In: Voorhees EM, Harman DK (eds) *TREC: experiment and evaluation in information retrieval*. MIT Press, Boston, chap 5, pp 99–121
- Robertson S, Hancock-Beaulieu M (1992) On the evaluation of IR systems. *Inf Process Manag* 28(4):457–466
- Sakai T (2006) Evaluating evaluation metrics based on the bootstrap. In: *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '06*. ACM, New York, pp 525–532
- Sakai T (2008a) Comparing metrics across TREC and NTCIR: the robustness to pool depth bias. In: *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval*, pp 691–692
- Sakai T (2008b) Comparing metrics across TREC and NTCIR: the robustness to system bias. In: *Proceedings of the 17th ACM conference on information and knowledge management*, pp 581–590
- Sakai T (2014) Metrics, statistics, tests. In: Ferro N (ed) *2013 PROMISE winter school: bridging between information retrieval and databases*. Lecture notes in computer science (LNCS), vol 8173. Springer, Heidelberg, pp 116–163
- Sakai T (2016) Statistical significance, power, and sample sizes: a systematic review of SIGIR and TOIS, 2006-2015. In: *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval, SIGIR '16*. ACM, New York, pp 5–14
- Sakai T, Kando N (2008) On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Inf Retr* 11:447–470
- Sanderson M (2010) Test collection based evaluation of information retrieval systems. *Found Trends Inf Retr* 4(4):247–375
- Sanderson M, Zobel J (2005) Information retrieval system evaluation: effort, sensitivity, and reliability. In: *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '05*. ACM, New York, pp 162–169
- Savoy J (1997) Statistical inference in retrieval effectiveness evaluation. *Inf Process Manag* 33(4):495–512
- Smucker MD, Allan J, Carterette B (2007) A comparison of statistical significance tests for information retrieval evaluation. In: *Proceedings of the sixteenth ACM conference on conference on information and knowledge management, CIKM '07*. ACM, New York, pp 623–632

- Soboroff I, Robertson S (2003) Building a filtering test collection for trec 2002. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '03. ACM, New York, pp 243–250
- Spärck Jones K (1974) Automatic indexing. *J Doc* 30:393–432
- Spärck Jones K (2001) Automatic language and information processing: rethinking evaluation. *Nat Lang Eng* 7(1):29–46
- Spärck Jones K, Bates RG (1977) Report on a design study for the 'IDEAL' information retrieval test collection. Tech. rep., Computer Laboratory, University of Cambridge. Available at <http://sigir.org/resources/museum/>
- Spärck Jones K, Van Rijsbergen C (1975) Report on the need for and provision for and 'IDEAL' information retrieval test collection. Tech. rep., Computer Laboratory, University of Cambridge. Available at <http://sigir.org/resources/museum/>
- Taube M (1965) A note on the pseudomathematics of relevance. *Am Doc* 16(2):69–72
- Tomlinson S, Hedin B (2011) Measuring effectiveness in the TREC legal track. In: Lupu M, Mayer K, Tait J, Trippe A (eds) Current challenges in patent information retrieval. The information retrieval series, vol 29. Springer, Berlin, pp 167–180
- Trotman A, Geva S, Kamps J, Lalmas M, Murdock V (2010) Current research in focused retrieval and result aggregation. *Inf Retr* 13(5):407–411
- Turpin AH, Hersh W (2001) Why batch and user evaluations do not give the same results. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '01, pp 225–231
- Van Rijsbergen C (1979) Evaluation, 2nd edn. Butterworths, London, chap 7
- Voorhees EM (2000) Variations in relevance judgments and the measurement of retrieval effectiveness. *Inf Process Process* 36:697–716
- Voorhees EM (2005) Question answering in TREC. In: Voorhees EM, Harman DK (eds) TREC: experiment and evaluation in information retrieval. MIT Press, Boston, chap 10, pp 233–257
- Voorhees EM (2014) The effect of sampling strategy on inferred measures. In: Proceedings of the 37th international ACM SIGIR conference on research &#38; development in information retrieval, SIGIR '14. ACM, New York, pp 1119–1122
- Voorhees EM, Buckley C (2002) The effect of topic set size on retrieval experiment error. In: Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '02. ACM, New York, pp 316–323
- Voorhees EM, Harman DK (2005) The Text REtrieval Conference. In: Voorhees EM, Harman DK (eds) TREC: experiment and evaluation in information retrieval. MIT Press, Boston, chap 1, pp 3–19
- Yilmaz E, Aslam JA (2008) Estimating average precision when judgments are incomplete. *Knowl Inf Syst* 16:173–211
- Yilmaz E, Kanoulas E, Aslam JA (2008) A simple and efficient sampling method for estimating AP and NDCG. In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, pp 603–610
- Zobel J (1998) How reliable are the results of large-scale information retrieval experiments? In: Croft WB, Moffat A, van Rijsbergen C, Wilkinson R, Zobel J (eds) Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, Melbourne, Australia. ACM Press, New York, pp 307–314

# How to Run an Evaluation Task



## With a Primary Focus on Ad Hoc Information Retrieval

Tetsuya Sakai

**Abstract** This chapter provides a general guideline for researchers who are planning to run a shared evaluation task for the first time, with a primary focus on simple *ad hoc Information Retrieval (IR)*. That is, it is assumed that we have a static target document collection and a set of test topics (i.e., search requests), where participating systems are required to produce a ranked list of documents for each topic. The chapter provides a step-by-step description of what a task organiser team is expected to do. Section 1 discusses how to define the evaluation task; Sect. 2 how to publicise it and why it is important. Section 3 describes how to design and build test collections, as well as how inter-assessor agreement can be quantified. Section 4 explains how the results submitted by participants can be evaluated; examples of tools for computing evaluation measures and conducting statistical significance tests are provided. Finally, Sect. 5 discusses how the fruits of running the task should be shared to the research community, how progress should be monitored, and how we may be able to improve the task design for the next round. N.B.: A prerequisite to running a successful task is that you have a good team of organisers who can collaborate effectively. Each team member should be well-motivated and committed to running the task. They should respond to emails in a timely manner and should be able to meet deadlines. Organisers should be well-organised!

## 1 Task Definition

This section discusses how a new task should be defined and proposed. If your proposal is going to be reviewed (for example, at *Conference and Labs of the Evaluation Forum (CLEF)*, *NII Testbeds and Community for Information access Research (NTCIR)*, or *Text REtrieval Conference (TREC)*), the proposal document should be composed very carefully: if it is not accepted, the task will not happen.

---

T. Sakai (✉)  
Waseda University, Tokyo, Japan  
e-mail: [tetsuyasakai@acm.org](mailto:tetsuyasakai@acm.org)



Even if you are planning to run it outside existing evaluation forums, you still need to have a clear motivation and a task design.

## ***1.1 Research Motivation, Research Questions***

Why do you want to run this new task? How will the outcomes of the task help the research community or the world in general? What research questions do you want to address as a task organiser, and what research questions do you want your participants to address? Write down your answers. If you are proposing a new task to (say) CLEF and are completely new to CLEF, I recommend that you contact the organisers of existing CLEF tasks/labs, and ask them to share their proposal documents with you. Learning from positive examples helps.

Novelty is important. For example, if you propose a task for CLEF but that task is almost the same as the one TREC ran few years ago, your proposal may not be accepted. It is okay to propose a new task that is closely related to existing tasks, but if that is the case, highlight the differences. What are the new angles and/or new approaches to the evaluation?

## ***1.2 Input and Output Specifications***

You are organising a task and want your participants to submit files, code, etc. To accomplish this, you need to clearly define the input to each participating system, and the expected output from them. Henceforth, whenever concrete examples seem to help, I shall draw a few from the NTCIR-12 *We Want Web (WWW)* Task in which I was recently involved (Luo et al. 2017), since this was a typical ad hoc IR task.

### **1.2.1 Input: Topics**

In ad hoc IR, a *topic* is a search request that expresses a specific information need. Historically, TREC composed topics by having different *fields*, such as `title`, `description`, and `narrative` (Harman 2005); participants utilised different (combinations of) fields as input to their IR systems. However, a simpler input format is also possible: for example, a test topic set could be provided to participants as a text file, where each line comprises a topic ID (e.g., “0008”) and a query string (e.g., “World Table Tennis Championships”).

Whatever the input file format you choose, you should announce it to the participants much earlier than the topic set release date, so that they can develop a suitable parser for the input file in advance.

### 1.2.2 Output: Runs

An evaluation task typically requires participating teams to process the input file (e.g., the topic set file) and to submit one or more *runs*. In the context of ad hoc IR, a run is a file that contains a ranked list of retrieved documents (actually, just the IDs of those documents) for each topic.

Figure 1 shows what an actual run file for ad hoc IR looks like. In this particular example, except for the first line in the file that concisely describes how the run was generated, the run conforms to the standard “TREC run format.” Each line has the following six fields:

1. Topic ID;
2. Dummy field (usually not used in any way);
3. Document ID, a.k.a. DOCNO;
4. Rank of the document according to the participating system;
5. Score of the document;
6. Name of the run, i.e., same as the file name (usually not used in any way, unless you need to monitor at the pooling stage (See Sect. 3.2) which runs contributed which documents).

Personally, I find the additional SYSDISC line quite useful when writing the task overview paper although this deviates from the standard TREC run format: we can easily extract the concise description of each run in participants’ own words, instead of having to read each participant’s paper very carefully (which you should!) and to formulate our own descriptions of the runs.

Whatever your preferred output format is, you should announce the requirements well before the run submission deadline so that participants can be prepared. Specificity is important. In particular, I recommend the following:

- Clarify how many runs you are allowing each participating team to submit. Budget constraints may force you to evaluate only a subset of the submitted runs.

```
$ head -5 Runs/RUCIR-E-NU-Base-1
<SYSDISC>Rank with traditional similarity features</SYSDISC>
0001 0 clueweb12-0309wb-37-05516 1 -1.03 RUCIR-E-NU-Base-1
0001 0 clueweb12-0403wb-78-00139 2 -1.08 RUCIR-E-NU-Base-1
0001 0 clueweb12-0012wb-31-00153 3 -1.21 RUCIR-E-NU-Base-1
0001 0 clueweb12-0210wb-30-33571 4 -1.22 RUCIR-E-NU-Base-1
$ tail -5 Runs/RUCIR-E-NU-Base-1
0100 0 clueweb12-0204wb-68-21006 96 -0.67 RUCIR-E-NU-Base-1
0100 0 clueweb12-1500tw-66-16405 97 -0.67 RUCIR-E-NU-Base-1
0100 0 clueweb12-1406wb-97-04314 98 -0.68 RUCIR-E-NU-Base-1
0100 0 clueweb12-0301wb-03-31258 99 -0.68 RUCIR-E-NU-Base-1
0100 0 clueweb12-0606wb-03-20426 100 -0.68 RUCIR-E-NU-Base-1
```

Fig. 1 Parts of an actual run file from an ad hoc web search task (Luo et al. 2017)

Hence, it would be good to ask the participants to prioritise their runs, e.g., “*Up to five runs can be submitted and they should be numbered 1–5 according to priority. It is possible that only the first few of your runs will be included in the pool.*”

- Establish a run file naming convention. It would be convenient if you ask each team to prefix their run names with their team names, and to include the priority number in the run names. For example, in Fig. 1, the run is from Team RUCIR, and its priority number is 1. Moreover, if the task accommodates several different run *types*, it would be convenient to have them clearly encoded in the run file names. For example, the substring E-NU-Base of the run file name RUCIR-E-NU-Base-1 carries a specific meaning: it is a run submitted to the English subtask (E); it did not utilise the query log data provided by the organisers (NU); it utilised the baseline run provided by the organisers (Base). Systematic naming of run files facilitates later analysis.
- Clarify whether *manual* runs are allowed: if there was some manual intervention in at least one step in the entire process of generating the run, however minor (e.g., manually correcting the spelling of the input query), that run is usually considered manual. Otherwise, runs are considered *automatic*. It is recommended to encode this distinction also in the aforementioned run type information.

### 1.3 *Timeline, Budgeting, Expected Outcomes*

Scheduling is also very important when planning a task. If you are running a task under an umbrella such as CLEF, NTCIR, TREC, etc., then there will be several constraints that you will have to take into account. For example, there will be deadlines for your task overview paper and for the participants’ papers. Under these constraints, you need to consider the following points at least:

- How much time are you giving the participants for developing and training their systems? For example, if you plan to provide training data to participants and you expect that they require 2 months for training, you need to release the training data at least 2 months prior to the *test period*: see below.
- By *test period*, I mean the time between the *test topic set release date* and the *run submission deadline*. If you are looking for practical IR methods that process the test topics automatically and efficiently, you might let the test period be 1 week, for example. However, it should be noted that tasks participants are probably not working on your task *full time*, and that setting a high bar in this way may mean you are letting some registered participants drop out. If you expect participants to try computationally expensive approaches, you might choose a longer test period, say, 1 month. In any case, your instructions should clearly state that the participants are not allowed to tune their systems with the test data.
- How much time do you need to create the gold data, i.e., the ground truth? In the case of IR based on *pooling* (See Sect. 3.2), building the gold data, i.e.,

relevance assessments, takes place *after* the run submission deadline. How many documents can a relevance assessor judge in an hour? How many relevance assessors can you hire, and how many assessors can you assign to each topic? If they work in parallel, how much time will you need to complete the labelling task? You should allow for a little buffer period for accidents that may happen: an assessor may get sick and have to be replaced; the server for your relevance assessment tool (See Sect. 3.3.1) may get struck by a lightning and break down. A little pessimism would be good.

If you have designed a completely new task, it would help to have a *dryrun* prior to the formal run period. For example, if you have 100 topics for the formal run (Why 100 topics? Let us discuss this in Sect. 3.1), you might use a separate set of 5 topics for the dryrun, and let participants and yourself “practice.” This will help you detect problems with the protocol (e.g., the topic set format, run file format, etc.) or the tools you plan to use for the evaluation. It would be good to be rid of these problems *before* spending a lot of resources on the formal run evaluation.

#### 1.4 Plans for Gold Data and Evaluation Measures

At the planning stage, you need to be clear about how you are going to create the gold standard data, i.e., relevance assessments in the case of ad hoc IR. If the amount of assessment work depends on how many runs you receive from the participants, there is no need to finalise the specifics at this point. For example, if you are conducting an ad hoc IR task and plan to create relevance assessments by pooling (See Sect. 3.2), it may not be possible to decide on the *pool depth* (i.e., the number of documents to scoop from the top of each run to form a pool of documents) until you have actually received the run files.

For ad hoc IR relevance assessments, it would be good to clearly define what your *relevance levels* are going to be. Early TREC tracks used binary relevance assessments and binary relevance measures such as precision, recall, and Average Precision (Buckley and Voorhees 2005; Harman 2005), but modern IR test collections accommodate *graded* relevance. For example, you might want to have *ordinal* relevance levels such as *highly relevant*, *relevant*, and *nonrelevant*.<sup>1</sup> In this case, you need to write down the definition of each relevance level. Relevance assessors are humans, and humans inevitably disagree with each other and even with themselves when judging documents. Hence, it is important for you to clearly define the relevance levels; it is your responsibility, not theirs. Relevance assessments may

---

<sup>1</sup>The difference between *highly relevant* and *relevant* is not necessarily equivalent to that between *relevant* and *nonrelevant*.

also be collected on an *interval* scale: for example, by asking assessors to choose from 1, 2, 3, and 4, where 4 represents the highest relevance.<sup>2</sup>

Depending on what you want to evaluate, you may reuse existing evaluation measures or invent one for yourself. In either case, it would help to conduct a preliminary study to investigate the properties of the evaluation measures. In particular, for designing a new test collection from a statistical point of view, it would be good to know how statistically *stable* your measures are: this will be discussed in Sect. 3.1. You should announce the primary evaluation measures that you plan to use as early as possible, if you expect participants to want to tune their systems according to those measures. This does not prevent you from trying alternative measures after the run submission. If your gold data is graded, then I recommend that you use evaluation measures that can handle graded relevance. For example, if you want a search engine that can rank highly relevant documents on top of less relevant ones, you should use evaluation measures that are based on that goal. Section 4.1 briefly describes how some popular (and not that popular) evaluation measures can be computed using a publicly available tool.

## 2 Publicity

The task you proposed to CLEF, NTCIR, or TREC has been accepted? Congratulations! This section discusses how to attract as many research groups as possible to the task and why you should try to do so.

### 2.1 Why Attract Many Research Groups?

Why would you want many research groups to participate in your task? Because in general, a large task means a large impact on the research community. For example, more people who participated in your task will publish papers related to your task. More people will then cite your overview paper, and then eventually new people will be attracted to your task.

It is not just about the number of participating groups. You need *good* participating groups. For example, if you run an IR task in which only a few groups with no prior experience in IR participated, that task is not likely to advance the state of the art. You also need *diverse* approaches from the participating groups. For example, if you run an IR task in which all of your participants take the same approach using the same off-the-shelf search engine, you may not discover much

---

<sup>2</sup>The assessors should be made aware that the difference between 1 and 2 is the same as that between 2 and 3, and so on.

from the task outcomes. In particular, if you plan to pool submitted runs to obtain relevance assessments, you really need good and diverse participating groups in order to build reliable document pools and thereby make your test collection as *reusable* as possible (See Sect. 3.2).

## 2.2 *How to Attract Many Research Groups*

You need to advertise your task. Create a website and social media accounts (e.g., Twitter and Facebook). Send out Call for Participation messages to relevant mailing lists several times well before your task begins. It is very important to disclose as many details about the tasks as possible at the early stage, as we have discussed in Sect. 1: it would be difficult for participants to participate in a task in which everything is “TBD.” Giving them details early on also means a higher chance for them to perform well in your task.

Be responsive to inquiries from prospective and registered participants. When providing answers to the inquiries, make sure that you are giving *all* participants the same information. Fairness is important in a shared task, even if it is not a competition. Speaking of fairness, if you are allowing yourself to participate in the task, clearly state in the official results which runs are organisers’ runs. Since organisers inevitably have access to more information than regular participants, some may choose to refrain from submitting their own runs. I recommend organisers to submit “organisers’ baseline runs” using standard approaches. This would be useful not only for relative comparisons of participants’ approaches but also for enhancing the quality of your document pools.

## 3 **Designing and Building Test Collections**

This section describes how IR test collections can be designed and constructed through a shared task, with specific examples from ad hoc IR. Section 3.1 describes statistical techniques for determining the topic set size. Section 3.2 discusses how pooling can be conducted based on the submitted runs. Finally, Sect. 3.3 describes how relevance assessments can be conducted, and how inter-assessor agreement can be checked.

### 3.1 Topic Set Sizes

It is common to have about 50 test topics for a task: this practice probably originates from TREC which was launched in the early 1990s (Harman 2005). Whereas, Sparck Jones and van Rijsbergen (1975) once argued:

- < 75 requests are of no real value,
- 250 requests are minimally acceptable,
- > 1000 requests are needed for some purposes.

Also, the TREC 2007 Million Query Track had about 1800 topics by employing statistically motivated techniques for efficient relevance assessments (Allan et al. 2008). For a more detailed discussion on the history of topic set sizes in IR, see Sakai (2018a).

So how many topics should you create for your new test collection? If you expect the future users of your new test collection to conduct statistical significance tests and you want them to reach reliable conclusions, it is possible to determine the topic set size based on *sample size design* (Nagata 2003). In the IR context, the approaches are called *topic set size design* (Sakai 2016).

The first thing that researchers hoping to build “statistically reliable” IR test collections should be aware of is that the required topic set size depends on how statistically stable your *evaluation measure* is. Here, the stability of a measure can be understood as its *population variance*, which we assume to be system-independent for the sake of convenience. To be more specific, we tentatively assume that the evaluation scores for the  $i$ -th system obeys  $N(\mu_i, \sigma^2)$ , where  $\mu_i$  is the population mean for the  $i$ -th system and  $\sigma^2$  is the common population variance.<sup>3</sup>

Suppose you have a small pilot data set and a few baseline IR systems from which you can obtain a  $n \times m$  topic-by-run matrix where each element is a score according to the evaluation measure of your choice. By treating this as a two-way *ANalysis Of VAriance (ANOVA)* (without replication) matrix, you can first compute a residual sum of squares  $S_{E2}$  after removing the between-system and between-topic sums of squares  $S_A$  and  $S_B$  from the total sum of squares  $S_T$ , and then obtain an unbiased estimate of  $\sigma^2$  as:

$$\hat{\sigma}^2 = V_{E2} = \frac{S_{E2}}{(m-1)(n-1)}. \quad (1)$$

Furthermore, if multiple topic-by-run matrices for the same evaluation measure are available,  $\hat{\sigma}^2$  can be computed as a *pooled variance* based on the residual variance from each matrix; see Sect. 5.2.

---

<sup>3</sup>Even if the evaluation scores violate the normality assumption, the *Central Limit Theorem* says that *mean* scores can be regarded as normality distributed, provided that the topic set size is sufficiently large (Sakai 2018a).

The larger  $\hat{\sigma}^2$  is, the larger your topic set will have to be in order to ensure reliable research conclusions. That is why it is recommended to study the evaluation measure of your choice in advance using some small pilot data from your dryrun or an existing test collection (even if that test collection was not designed for your new task—you need to start somewhere). Moreover, even if you have decided on your primary evaluation measure, you should be aware that  $\hat{\sigma}^2$  depends also on the *pool depth* (or, more generally, how many relevant documents you have identified in your test collection) and the *measurement depth* (i.e., how many top documents you are going to look at to compute your favourite evaluation measure). For example, if you are interested in evaluating the first *Search Engine Result Page (SERP)* in a web search task, your measurement depth may be 10; evaluation measures at cutoff 10 generally have higher variances than those at (say) 1000, since the former rely on considerably fewer data points. Shallow pools (i.e., small pool depths) also imply fewer data points and therefore higher variances.

The tradeoff between the topic set size and the pool depth is especially important as it directly affects the relevance assessment cost. Several studies have shown that, from a statistical point of view, it is better to have a large topic set with shallow pools than to have a small topic set with deep pools (e.g., Carterette et al. 2008b; Webber et al. 2008). For example, the topic set size design analysis of Sakai (2016) based on the TREC 2003 and 2004 Robust track data with Average Precision suggests that a test collection with depth-10 pools for 101 topics is statistically equivalent (for the purpose of conducting paired *t*-tests) to another with depth-100 pools for 76 topics: the former requires less than 10,000 relevance assessments, which is only about 17.5% of the latter case.

The above argument appears to suggest that we should always go for a large topic set with *very* shallow pools. However, note that a shallow pool depth means that your relevance assessments will be highly *incomplete*: that is, there will be many unidentified relevant documents in the target document collection (Buckley and Voorhees 2004, 2005; Sakai 2007; Voorhees 2002). This may limit the *reusability* of your test collection: for example, suppose that after you have released your test collection to the research community, a new research group evaluates their novel system with that collection. Their system is so novel that it manages to retrieve some unidentified relevant documents, but none of the relevant documents that you have identified: their effectiveness score will then be zero. While this is an extreme case, systems that did not participate in the pooling process are likely to be underestimated relative to those that did. In summary: have a large topic set with pools that are “not too deep.” Also, if your pool depth is larger than the official measurement depth (e.g., 30 vs. 10), that is more reassuring than if it is smaller.<sup>4</sup>

---

<sup>4</sup>It should be noted that the ad hoc tracks of TREC typically used depth-100 pools while the measurement depth was 1000, which means that many of the retrieved documents remained unjudged. This may be a problem (Zobel 1998), but remedies exist: top-heavy evaluation measures (i.e., those that rely primarily on the top 100 rather than top 1000), or even evaluation measures especially designed for incomplete relevance assessments (e.g., Buckley and Voorhees 2004; Yilmaz and Aslam 2006; Sakai 2007).



Suppose you have obtained a variance estimate  $\hat{\sigma}^2$  for your favourite evaluation measure using Eq. (1) with some pilot data. Section 3.1.1 describes three simple Microsoft Excel tools for determining the topic set size for your new test collection based on statistical power; Sect. 3.1.2 describes two tools for a similar purpose based on confidence interval (CI) widths. The five tools are available from my website.<sup>5</sup> Three of these tools are based on standard statistical techniques for handling *unpaired* data, and they require  $\hat{\sigma}^2$  as a direct input that is fed to the tools. Whereas, the other two tools are based on the *t*-test and the CIs for *paired* data, respectively, and require an estimate of the variance ( $\sigma_d^2 = \sigma_X^2 + \sigma_Y^2$ ) of the *difference* between two systems *X* and *Y*, where  $\sigma_X^2$  and  $\sigma_Y^2$  are the population variances for the two systems (Sakai 2018b). Hereafter, we follow Sakai (2016, 2018a,b) and simply consider cases where we let  $\hat{\sigma}_d^2 = 2\hat{\sigma}^2$  in order to utilise these two tools based on paired data; that is, we assume *homoscedasticity* (i.e., equal variances) even with paired data.

It should be noted here that the topic set size design tools rely on statistical power and CIs that regard the topic set as a *random sample* from the population. That is, the basic assumption is that you have (say) a large query log and you draw a query at random *n* times independently to obtain a topic set of size *n*. However, IR topic sets are rarely random samples, since test collection builders often hand-pick queries in the hope of testing some specific features of IR systems. Section 4.2 includes a discussion on a randomisation-based approach to significance testing, which does not rely on the random sampling assumption.

### 3.1.1 Power-Based Topic Set Size Design

Three Excel tools are available for topic set size design based on a statistical power requirement: `samplesizeTTEST2.xlsx` which is based on the paired *t*-test, `samplesize2SAMPLET.xlsx` which is based on the two-sample (i.e., unpaired) *t*-test, and `samplesizeANOVA2.xlsx` which is based on one-way ANOVA. While the latter two tools require a variance estimate  $\hat{\sigma}^2$  as an input, recall that `samplesizeTTEST2.xlsx` requires  $\hat{\sigma}_d^2$ , the estimated variance of the score *differences*. If the researcher is interested in ensuring a high statistical power for comparing any  $m = 2$  systems in terms of a particular evaluation measure (regardless of whether the data are paired or not), I recommend the use of `samplesizeANOVA2.xlsx` to obtain the required topic set size, as it can provide the tightest topic size estimates while avoiding the use of  $\hat{\sigma}_d^2$ . While the two-sample *t*-test and one-way ANOVA for  $m = 2$  systems are theoretically equivalent, the two tools yield slightly different results due to different approximations involved in the power calculations (Sakai 2018a,b).

---

<sup>5</sup><http://sakailab.com/download/>.

`sampleSizeANOVA2.xlsx` requires the following as input:

- $\alpha$  Probability of Type I Error (i.e., concluding that the population means are different even though they are not);
- $\beta$  Probability of Type II Error (i.e., concluding that the population means are equal even though they are not);
- $m$  Number of systems that are compared;
- $minD$  *Minimum detectable range*. That is, whenever the true difference between the best and the worst among the  $m$  systems is  $minD$  or larger, we want to guarantee  $(1 - \beta)\%$  statistical power. Note that when  $m = 2$ ,  $minD$  simply represents the difference between the two systems that are being compared.
- $\hat{\sigma}^2$  Variance estimate for a particular evaluation measure (See Eq. (1)).

For example, suppose you are interested in the statistical power for comparing  $m = 10$  systems with ANOVA at  $\alpha = 0.05$  with an evaluation measure whose estimated variance is  $\hat{\sigma}^2 = 0.1$ . If you want to guarantee 80% power whenever the difference between the best and the worst systems is 0.1 or larger, choose the sheet for  $(\alpha, \beta) = (0.05, 0.20)$  in `sampleSizeANOVA2.xlsx` and enter  $(m, minD, \hat{\sigma}^2) = (10, 0.10, 0.10)$ : you will obtain 312 as the recommended topic set size.

### 3.1.2 CI-Based Topic Set Size Design

Two Excel tools are available for topic set size design based on a desired cap on the CI width for the difference between two systems: `sampleSizeCI2.xlsx` is for paired-data CIs, while `sampleSize2SAMPLECI.xlsx` is for unpaired-data CIs. Again, while the latter requires  $\hat{\sigma}^2$ , the former requires  $\hat{\sigma}_d^2$ , which is harder to estimate. Hence, if the researcher is interested in topic set size design based on CI widths, I recommend the use of `sampleSize2SAMPLECI.xlsx` regardless of whether the data are paired or not: the topic set sizes thus obtained are always large enough for paired-data cases as well.<sup>6</sup>

`sampleSize2SAMPLECI.xlsx` requires the following as input:

- $\alpha$  Type I Error probability. This is usually set to 0.05 as we usually want to discuss 95% confidence intervals.
- $\delta$  An upperbound for the width of the confidence interval for the difference between any system pair. That is, you want the confidence interval to be no wider than  $\delta$ .
- $\hat{\sigma}^2$  Variance estimate for a particular evaluation measure.

For example, suppose that you want the width of a 95% confidence interval for the difference between any system pair to be no larger than 0.1, for an evaluation measure whose estimated variance is  $\hat{\sigma}^2 = 0.1$ . Enter  $(\alpha, \delta, \hat{\sigma}^2) = (0.05, 0.10, 0.10)$  to

---

<sup>6</sup>Unlike the earlier CI-based tool described in Sakai (2016), these CI-based tools can handle large topic set sizes without any problems; see Sakai (2018b) for details.

`samplesize2SAMPLECI.xlsx`: you will obtain 309 as the recommended topic set size.

## 3.2 Pooling

Sparck Jones and van Rijsbergen (1975) discussed the idea of pooling for situations where it is difficult to obtain exhaustive relevance assessments:

Ideally these should be exhaustive. But if not some attempt should be made to carry out independent searches using any available information and device, to obtain a pooled output for more broadly based relevance judgements than may be obtained only with simple user evaluation of standard search output.

This idea was implemented on a large-scale (compared to previous studies at the time) at TREC, which was launched in the early 1990s.

A simple pooling method works as follows: for a particular topic, let  $D_i(k)$  be the set of top  $k$  documents from the  $i$ -th submitted run; let  $s$  be the number of submitted runs. Then the depth- $k$  pool for that topic is given by

$$\bigcup_{i=1}^s D_i(k). \quad (2)$$

Note that  $|\bigcup_{i=1}^s D_i(k)| \leq sk$ . That is, the pool size, or the number of documents to be judged, is bounded above by  $sk$ ; hence, if depth  $k$  is going to be applied to all of the  $n$  topics, the total number of documents to be judged is bounded above by  $nsk$ , which will enable you to estimate the relevance assessment cost.

Let us run through examples with actual data: download `CLEF20sakai.tar.gz` from my website.<sup>7</sup> It contains three actual runs: the RMIT run and the RUCIR run contain a total of 10,000 documents (excluding the aforementioned SYDESC line) or 100 documents per topic; the THUIR run contains a total of 100,000 documents (excluding the aforementioned SYDESC line) or 1000 documents per topic. Next, you can download the NTCIREVAL toolkit<sup>8</sup>: we shall use this for computing evaluation measures in Sect. 4 as well. Here, we are only using a simple script called `TRECSplitruns` from NTCIREVAL, which splits the raw submitted run files (in TREC-like format) into *per-topic res* (i.e., result) files.<sup>9</sup> In this section, we shall use the `res` files to create pool files for relevance assessments.

`CLEF20sakai.tar.gz` also contains a file called `Etidlist` which is just a list of topic IDs (100 of them). Figure 2 shows how `TRECSplitruns` can be used

<sup>7</sup><http://sakailab.com/download/>.

<sup>8</sup><http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html>.

<sup>9</sup>`TRECSplitruns` does not change the document order for each topic in the original run file in any way; it disregards the ranks and scores in the run file. See `README` files for more details.

```

$ ls Runs/*
Runs/RMIT-E-NU-Own-1  Runs/RUCIR-E-NU-Base-1  Runs/THUIR-E-PU-Base-3
$ ls Runs/* | TRECSplitruns Etidlist 100
created 0001/0001.RMIT-E-NU-Own-1.res
created 0002/0002.RMIT-E-NU-Own-1.res
created 0003/0003.RMIT-E-NU-Own-1.res
:
created 0098/0098.THUIR-E-PU-Base-3.res
created 0099/0099.THUIR-E-PU-Base-3.res
created 0100/0100.THUIR-E-PU-Base-3.res

```

**Fig. 2** Splitting the run files into per-topic `res` files using `TRECSplitruns` from the `NTCIREVAL` toolkit

```

$ cat runlist
RMIT-E-NU-Own-1
RUCIR-E-NU-Base-1
THUIR-E-PU-Base-3
$ cat Etidlist | Sortedpool-foreach-topic runlist 100
created 0001.runlist.pd100.pool
created 0001.runlist.pd100.sortpool
created 0002.runlist.pd100.pool
created 0002.runlist.pd100.sortpool
:
created 0100.runlist.pd100.pool
created 0100.runlist.pd100.sortpool

```

**Fig. 3** Creating pool files using `Sortedpool-foreach-topic` from the `NTCIRPOOL` toolkit

to create per-topic `res` files for each run: it can be observed that the first argument is the list of topic IDs. As for the second argument, this is the cutoff applied when creating the `res` files: in this case, only the top 100 documents are kept in each `res` file (even though the original THUIR run file contains 1000 documents per topic). You can verify for yourself that we have a total of  $3 \text{ systems} \times 100 \text{ topics} \times 100 \text{ documents} = 30,000 \text{ documents}$  (including duplicates) in the `res` files.

Now you can download the `NTCIRPOOL` toolkit.<sup>10</sup> Let us try using a script called `Sortedpool-foreach-topic` to create a pool file for each topic. Figure 3 shows how it can be used: note that the file `runlist` contains the three run file names; the second argument specifies the pool depth  $k$ . Note that since we already truncated the raw run files at cutoff 100 in Fig. 2, setting the pool depth  $k$  to a value larger than 100 in Fig. 3 will have the same effect as setting it to 100.

`Sortedpool-foreach-topic` actually creates two different pool files for each topic. Here, the `pool` file contains the IDs of the pooled documents sorted alphabetically. Whereas, the `sortpool` file contains three additional fields and the

<sup>10</sup><http://research.nii.ac.jp/ntcir/tools/ntcirpool-en.html>.

same set of documents are arranged in decreasing order of the value of the fourth field. The meanings of these fields are as follows:

Second field Number of runs that returned this document (*nruns*);

Third field Rank of this document in each run summed across the *nruns* (*sumranks*);

Fourth field Document sort key defined as  $nruns + \frac{1}{sumranks}$ .

The idea is that documents that have been returned by many systems at high ranks are more likely to be relevant than others (Sakai and Lin 2010). While TREC traditionally presents pooled documents sorted by the document IDs (Harman 2005) (as in the `pool` files) to relevance assessors, several tasks of NTCIR (e.g., *Advanced CrossLingual Information Access (ACLIA)*, *INTENT*, *Short Text Conversation (STC)*, and *WWW* tasks) have used the `sortpool` files instead. The former approach tries to randomise the document presentation order, while the latter is designed to let the assessor form an idea as to what constitutes a relevant document at an early stage of the judging process in an environment where similar documents are presented relatively close to each other.<sup>11</sup>

Finally, you can verify that the total number of lines in the `pool` (or `sortpool`) files is 23,960, which is indeed smaller than 30,000 due to the overlap of retrieved documents across the three runs.

### 3.3 Relevance Assessments and Relevance Levels

Section 3.3.1 discusses an example of a relevance assessment tool; Sect. 3.3.2 describes measures for checking inter-assessor agreement; finally, Sect. 3.3.3 touches upon a few novel approaches to obtaining relevance assessments.

#### 3.3.1 Relevance Assessment Tool

Once you have prepared the `pool` files, you need to assign them to relevance assessors. You need to provide a relevance assessment tool to each assessor so that they can process the documents efficiently and reliably; the tool should also let you monitor the progress of each assessor. For each topic, the pooled documents should be loaded onto the interface. Figure 4 shows a relevance assessment tool used for the NTCIR-13 WWW task (Luo et al. 2017): it has a typical user interface,

---

<sup>11</sup>Several researchers have remarked to me over the years that the “sorted pool” approach introduces a rank bias to the judgements. While this may be true, I believe that there are also merits, which I hope to demonstrate in a future study.

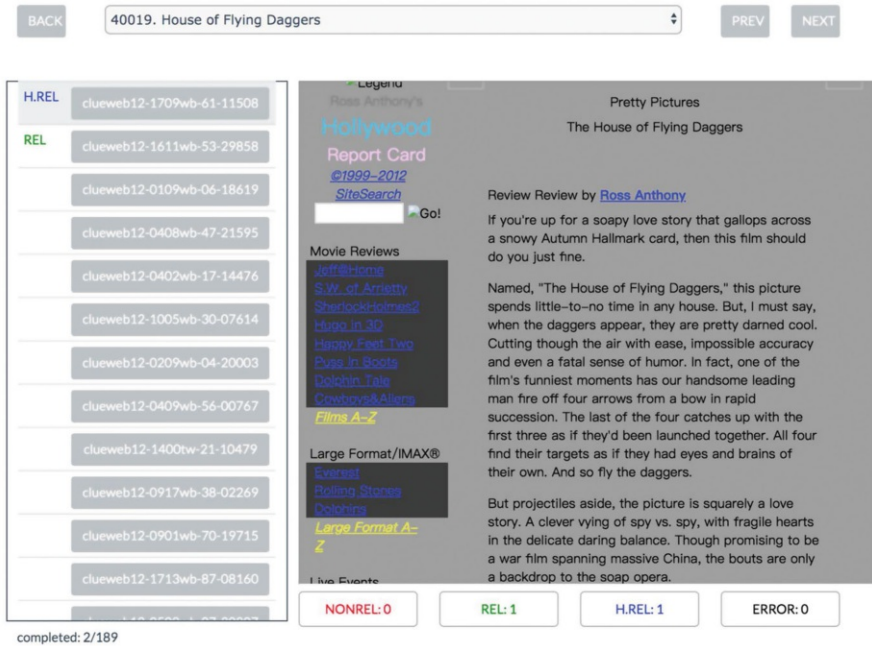


Fig. 4 A typical interface of a relevance assessment tool

with a document viewer panel on the right, relevance level buttons beneath it, and a document list panel on the left.<sup>12</sup>

As we have discussed in Sect. 1.4, your relevance assessors should be given clear instructions: the instructions should be given in a written form, and should be the same across all assessors.

### 3.3.2 Inter-Assessor Agreement Measures

It would be good to assign multiple assessors to at least a subset of your topic set, to check inter-assessor agreement, which is an indication of the reliability of your gold data. While *raw agreement*, a.k.a. *overlap*, has been used for quantifying inter-assessor agreement (e.g., Voorhees 2000) it should be noted that this measure does not take *chance agreement* into account. In this section, I first describe *Cohen’s  $\kappa$*  (Cohen 1960; Bailey et al. 2008) and *Cohen’s weighted  $\kappa$*  (Cohen 1968; Sakai 2015) for comparing the agreement of two assessors for nominal and ordinal scales.

<sup>12</sup>This tool, called *PLY*, was developed by Peng Xiao, Lingtao Li, and Yimeng Fang at the Sakai Laboratory, Waseda University.

**Table 1** An example for computing Cohen’s  $\kappa$ 

|          |       | (I) Observed      |      |       | (II) Expected     |      |       |
|----------|-------|-------------------|------|-------|-------------------|------|-------|
|          |       | Assessor <i>B</i> |      | Total | Assessor <i>B</i> |      | Total |
|          |       | rel               | nonr |       | rel               | nonr |       |
| Assessor | rel   | 50                | 30   | 80    | 48                | 32   | 80    |
|          | nonr  | 10                | 10   | 20    | 12                | 8    | 20    |
|          | Total | 60                | 40   | 100   | 60                | 40   | 100   |

Then, I describe Krippendorff’s  $\alpha$ -agreement for coding ( ${}_c\alpha$ ) (Krippendorff 2013),<sup>13</sup> which to my knowledge is the most versatile and robust agreement measure, in that it can handle nominal, ordinal, interval, and ratio categories for any number of assessors, and even cases where some labels are missing.

Fleiss’  $\kappa$  (Fleiss 1971) and its variant known as Randolph’s  $\kappa_{free}$  (Randolph 2005) were designed for measuring the agreement across more than two assessors, but they can only handle *nominal* categories, e.g., when three assessors assign either CLEF, NTCIR, or TREC to each research paper. Krippendorff’s  ${}_c\alpha$  is clearly more versatile. On the other hand,  ${}_c\alpha$  measures the reliability of data solely in terms of whether the labels agree with one another, and disregards which labels come from which assessors. For the purpose of detecting outlier assessors, I still find it useful to compute Cohen’s (weighted)  $\kappa$  for every pair of assessors per topic (Sakai 2017a).

### Cohen’s $\kappa$ and Weighted $\kappa$

Below, I borrow examples from Sakai (2015) to explain Cohen’s  $\kappa$  and weighted  $\kappa$  for measuring the agreement between two assessors beyond chance.

Table 1 shows a situation where Cohen’s  $\kappa$  can be applied: we have two (nominal) categories, *rel* (relevant) and *nonr* (nonrelevant).<sup>14</sup> Let  $O_{i\bullet}$  be the number of documents labelled as Category  $i$  by Assessor  $A$  and  $O_{\bullet j}$  be the number of documents labelled as Category  $j$  by Assessor  $B$ ; Let  $O_{ij}$  be the number of documents labelled as Category  $i$  by Assessor  $A$  and as Category  $j$  by Assessor  $B$ . Let  $N$  be the total number of labelled documents. The raw agreement is given by:

$$P_o = \frac{\sum_i O_{ii}}{N}. \quad (3)$$

From Table 1(I),  $P_o = (50 + 10)/100 = 60\%$ , but this high number may just reflect the fact that both assessors tend to say *rel* more often than *nonr*. Hence

<sup>13</sup>In the context of IR evaluation, *coding* can be interpreted as “assigning a relevance level to each document.”

<sup>14</sup>Note that Cohen’s  $\kappa$  is applicable to more than two nominal categories, for a pair of assessors.

**Table 2** An example for computing Cohen’s weighted  $\kappa$  with linear weights

|               |        | (I) Observed      |     |      |       | (II) Expected     |      |      |       | (III) Weights     |     |      |
|---------------|--------|-------------------|-----|------|-------|-------------------|------|------|-------|-------------------|-----|------|
|               |        | Assessor <i>B</i> |     |      |       | Assessor <i>B</i> |      |      |       | Assessor <i>B</i> |     |      |
|               |        | h. rel            | rel | nonr | Total | h. rel            | rel  | nonr | Total | h. rel            | rel | nonr |
| Assessor<br>A | h. rel | 15                | 6   | 6    | 27    | 5.4               | 8.1  | 13.5 | 27    | 0                 | 1   | 2    |
|               | rel    | 3                 | 19  | 4    | 26    | 5.2               | 7.8  | 13   | 26    | 1                 | 0   | 1    |
|               | nonr   | 2                 | 5   | 40   | 47    | 9.4               | 14.1 | 23.5 | 47    | 2                 | 1   | 0    |
|               | Total  | 20                | 30  | 50   | 100   | 20                | 30   | 50   | 100   |                   |     |      |

Table 1(II) computes the expected agreement when the two sets of assessments are *independent*. That is, the four cells in Section (II) are given by:

$$C_{ij} = \frac{O_i \cdot O_j}{N} . \tag{4}$$

The agreement in this hypothetical situation is:

$$P_c = \frac{\sum_i C_{ii}}{N} . \tag{5}$$

From Table 1(II),  $P_c = (48 + 8)/100 = 56\%$ . Thus, the overlap would be over 50% even if the two sets of assessments are independent.

Cohen’s  $\kappa$  is a normalised measure of the observed agreement that goes beyond chance:

$$\kappa = \frac{P_o - P_c}{1 - P_c} = \frac{\sum_i O_{ii} - \sum_i C_{ii}}{N - \sum_i C_{ii}} . \tag{6}$$

Its range is  $[-1, 1]$ . To construct a  $100(1 - \alpha)\%$  confidence interval for  $\kappa$ , the following margin of error can be computed:

$$MOE = z_{\alpha/2} \sqrt{\frac{P_o(1 - P_o)}{N(1 - P_c)^2}} , \tag{7}$$

where  $z_p$  is the upper  $100P\%$   $z$ -value.<sup>15</sup> The reader should verify that for the example given in Table 1,  $\kappa = 0.0909$ , 95%CI  $[-0.1273, 0.3091]$ .

Table 2 shows a situation where Cohen’s *weighted*  $\kappa$  can be applied. This time, we have graded relevance data, where the labels h. rel, rel, nonrel can be regarded as ordinal (i.e., h. rel > rel > nonrel). Weighted  $\kappa$  is a generalisation of the aforementioned  $\kappa$ , although we use seemingly different formulations here. As before, we have  $O_{ij}$  and  $C_{ij}$  in Parts (I) and (II) of Table 2, respectively; what is

<sup>15</sup>NORM.S.INV(1 - P) in Microsoft Excel.



**Table 3** The value-by-unit (i.e., label-by-document) matrix

| Units:  | 1   | ...            | $u$ | ...            | $N$ |                |                      |
|---------|-----|----------------|-----|----------------|-----|----------------|----------------------|
| Values: | 1   | $n_{11}$       | ... | $n_{u1}$       | ... | $n_{N1}$       | $n_{\bullet 1}$      |
|         | :   | :              |     | :              |     | :              | :                    |
|         | :   | :              |     | :              |     | :              | :                    |
|         | $i$ | $n_{1i}$       | ... | $n_{ui}$       | ... | $n_{Ni}$       | $n_{\bullet i}$      |
|         | :   | :              |     | :              |     | :              | :                    |
|         | :   | :              |     | :              |     | :              | :                    |
|         | $v$ | $n_{1v}$       | ... | $n_{uv}$       | ... | $n_{Nv}$       | $n_{\bullet v}$      |
| Totals: |     | $n_{1\bullet}$ | ... | $n_{u\bullet}$ | ... | $n_{N\bullet}$ | $n_{\bullet\bullet}$ |

new is that we now have Part (III) which defines the *weights* for each combination of disagreements. In this particular example, each `rel-nonrel` or `h.rel-rel` disagreement weighs 1 point; while each `h.rel-nonrel` disagreement weighs 2 points.<sup>16</sup> Let  $W_{ij}$  denote these weights.

Weighted  $\kappa$  is given by:

$$\kappa = 1 - \frac{Q_o}{Q_c} = 1 - \frac{\sum_i \sum_j W_{ij} O_{ij}}{\sum_i \sum_j W_{ij} C_{ij}}, \quad (8)$$

where  $Q_o$ ,  $Q_c$  represent the observed and expected chance *disagreements* rather than agreements.<sup>17</sup> Weighted  $\kappa$  reduces to the original  $\kappa$  when  $W_{ii} = 0$  for all  $i$  and  $W_{ij} = 1$  for all  $(i, j)$  s.t.  $i \neq j$ . To compute a  $100(1 - \alpha)\%$  confidence interval, the margin of error can be computed as follows:

$$MOE = z_{\alpha/2} \sqrt{\frac{R_o - Q_o^2}{N Q_c^2}}, \quad R_o = \frac{\sum_i \sum_j W_{ij}^2 O_{ij}}{N}. \quad (9)$$

The reader should verify that for the above example,  $\kappa = 0.6056$ , 95%CI [0.4646, 0.7465].

The R library `irr` mentioned below contains a function for computing Cohen's (weighted)  $\kappa$ , called `kappa2`.

### Krippendorff's $c_\alpha$

Table 3 shows a generic form of the data that can be handled by Krippendorff's  $c_\alpha$ . For us, "units" are documents that are judged by multiple assessors (thus there

<sup>16</sup>This gives us a *linear-weighted* kappa. Another popular variant is a *quadratic-weighted* kappa, where, for example, each `h.rel-nonrel` disagreement weighs  $2^2 = 4$  points to heavily penalise the mismatch.

<sup>17</sup>If  $P_o$ ,  $P_c$  are agreements and  $Q_o$ ,  $Q_c$  are disagreements, then clearly  $P_o = 1 - Q_o$  and  $P_c = 1 - Q_c$ . Hence,  $(P_o - P_c)/(1 - P_c) = (1 - Q_o - 1 + Q_c)/(1 - 1 + Q_c) = 1 - Q_o/Q_c$ .

are  $N$  documents); “values” are the possible relevance labels that the assessors can choose from, e.g., nonrelevant vs. relevant (nominal), nonrelevant vs. relevant vs. perfect (ordinal), 1 vs. 2 vs. 3 vs. 4 (interval, if, for example, the difference between 2 and 1 is equivalent to that between 4 and 3). Whatever the case, let the number of possible labels be  $v$ . In the table cells,  $n_{ui}$  denotes the number of assessors who labelled document  $u$  with value  $i$  (e.g., relevant). Note that Krippendorff’s  $c\alpha$  already disregards which labels come from which assessors. In the Totals row,  $n_{u\bullet} = \sum_i n_{ui}$ , i.e., the number of labels assigned to document  $u$ . If there are  $a$  assessors,  $n_{u\bullet} \leq a$  holds, since some assessments may be missing. As for  $n_{\bullet i}$  in the rightmost column, it is defined as  $n_{\bullet i} = \sum_{u|n_{u\bullet} \geq 2} n_{ui}$ . That is, documents that received only one label (i.e.,  $u$ ’s s.t.  $n_{u\bullet} = 1$ ) are excluded, because this label cannot be *paired* with another label for the same document and hence cannot tell us anything about the reliability of the data. Finally,  $n_{\bullet\bullet} = \sum_{u|n_{u\bullet} \geq 2} \sum_i n_{ui}$  is the total number of pairable values in the data; clearly,  $n_{\bullet\bullet} \leq aN$  holds.

The next step is to construct the matrix of *observed coincidences* and that of *expected coincidences* from the value-by-unit matrix. A coincidence matrix is a symmetrical  $v \times v$  matrix, where  $v$  is the number of possible label values. Table 4 shows the coincidence matrices in their generic forms. The cells are defined as follows. For each cell  $(i, j)$  where  $i \neq j$ ,

$$o_{ij} = \sum_u \frac{n_{ui}n_{uj}}{n_{u\bullet} - 1}, \quad e_{ij} = \frac{n_{\bullet i}n_{\bullet j}}{n_{\bullet\bullet} - 1}. \tag{10}$$

Whereas, for each cell  $(i, i)$  in the diagonal,

$$o_{ii} = \sum_u \frac{n_{ui}(n_{ui} - 1)}{n_{u\bullet} - 1}, \quad e_{ii} = \frac{n_{\bullet i}(n_{\bullet i} - 1)}{n_{\bullet\bullet} - 1}. \tag{11}$$

The general form of Krippendorff’s  $c\alpha$  is given by:

$$c\alpha_{metric} = 1 - \frac{D_o}{D_e} = 1 - \frac{\sum_i \sum_{j>i} o_{ij} \quad metric \delta_{ij}^2}{\sum_i \sum_{j>i} e_{ij} \quad metric \delta_{ij}^2} \tag{12}$$

**Table 4** Coincidence matrices: observed (left) and expected (right)

|         |                 |     |                 |     |                 |                      |
|---------|-----------------|-----|-----------------|-----|-----------------|----------------------|
| Values: | 1               | ... | $j$             | ... | $v$             |                      |
| 1       | $o_{11}$        | ... | $o_{1j}$        | ... | $o_{1v}$        | $n_{1\bullet}$       |
| :       | :               |     | :               |     | :               | :                    |
| :       | :               |     | :               |     | :               | :                    |
| $i$     | $o_{i1}$        | ... | $o_{ij}$        | ... | $o_{iv}$        | $n_{i\bullet}$       |
| :       | :               |     | :               |     | :               | :                    |
| :       | :               |     | :               |     | :               | :                    |
| $v$     | $o_{v1}$        | ... | $o_{vj}$        | ... | $o_{vv}$        | $n_{v\bullet}$       |
|         | $n_{\bullet 1}$ | ... | $n_{\bullet j}$ | ... | $n_{\bullet v}$ | $n_{\bullet\bullet}$ |

|         |                 |     |                 |     |                 |                      |
|---------|-----------------|-----|-----------------|-----|-----------------|----------------------|
| Values: | 1               | ... | $j$             | ... | $v$             |                      |
| 1       | $e_{11}$        | ... | $e_{1j}$        | ... | $e_{1v}$        | $n_{1\bullet}$       |
| :       | :               |     | :               |     | :               | :                    |
| :       | :               |     | :               |     | :               | :                    |
| $i$     | $e_{i1}$        | ... | $e_{ij}$        | ... | $e_{iv}$        | $n_{i\bullet}$       |
| :       | :               |     | :               |     | :               | :                    |
| :       | :               |     | :               |     | :               | :                    |
| $v$     | $e_{v1}$        | ... | $e_{vj}$        | ... | $e_{vv}$        | $n_{v\bullet}$       |
| Values: | $n_{\bullet 1}$ | ... | $n_{\bullet j}$ | ... | $n_{\bullet v}$ | $n_{\bullet\bullet}$ |

where  $metric \in \{nominal, ordinal, interval, ratio\}$ ;  $D_o$  is a measure of the observed disagreement and  $D_e$  is a measure of the disagreement that can be expected when chance prevails. For perfect agreement,  ${}_c\alpha = 1$ ; when observed and expected disagreements are equal,  ${}_c\alpha = 0$ ; The measure may become negative for small sample sizes and systematic disagreements.

The *difference function*  $metric\delta_{ij}^2$  in Eq. (12) depends on whether  $metric$  is on a nominal, ordinal, interval, or ratio scale. For nominal data, let  $nominal\delta_{ij}^2 = 1$  iff  $i \neq j$ , and let  $nominal\delta_{ii}^2 = 0$  for all  $i$ . For other cases:

$$ordinal\delta_{ij}^2 = \left( \sum_{k=i}^j n_{\bullet k} - \frac{n_{\bullet i} + n_{\bullet j}}{2} \right)^2, \quad (13)$$

$$interval\delta_{ij}^2 = (i - j)^2, \quad (14)$$

$$ratio\delta_{ij}^2 = \left( \frac{i - j}{i + j} \right)^2. \quad (15)$$

Figure 5 shows an example of computing  ${}_c\alpha$  using R. To use the `kripp.alpha` function, note that the `irr` library must be installed first. The matrix represents raw data from four relevance assessors who independently judged 12 documents.<sup>18</sup> Here, the possible relevance labels are represented as integers 1–5. Calling `kripp.alpha` without the second argument means treating the matrix as nominal data; if the numbers are treated as ordinal data, it can be observed that  ${}_c\alpha = 0.815$ ; if they are treated as interval data,  ${}_c\alpha = 0.849$ . Make sure you use the appropriate difference function.

Table 5 shows an instance of a label-by-document matrix (see Table 3), constructed from the raw assessor-by-document matrix shown in Fig. 5.<sup>19</sup> Recall that  ${}_c\alpha$  ignores the label for Document 12: by definition,  $n_{\bullet 3} = 10$ , not 11;  $n_{\bullet\bullet} = 40$ , not 41.

Table 6 shows how the observed and expected coincidence matrices can be computed from Table 5.<sup>20</sup> Recall Eqs. (10) and (11).

Table 7 shows the difference functions  $ordinal\delta_{ij}^2$  and  $interval\delta_{ij}^2$  for the data in Table 5. Recall Eqs. (13) and (14): only  $ordinal\delta_{ij}^2$  depends on the values of  $n_{\bullet i}$ .

Krippendorff also discusses computing bootstrap confidence intervals for  ${}_c\alpha$ : the reader is referred to his book for more details (Krippendorff 2013).

<sup>18</sup>This matrix is equivalent to the  $4 \times 12$  nominal data matrix described in Chapter 12 of Krippendorff (2013).

<sup>19</sup>Adapted from Chapter 12 of Krippendorff (2013).

<sup>20</sup>Adapted from Chapter 12 of Krippendorff (2013), but corrects his typo for the cell (5, 1) in the expected coincidences matrix.

```

> a4d12
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
[1,]    1    2    3    3    2    1    4    1    2    NA    NA    NA
[2,]    1    2    3    3    2    2    4    1    2    5    NA    NA
[3,]   NA    3    3    3    2    3    4    2    2    5    1    3
[4,]    1    2    3    3    2    4    4    1    2    5    1    NA
> kripp.alpha(a4d12)
Krippendorff's alpha

Subjects = 12
Raters = 4
alpha = 0.743
> kripp.alpha(a4d12, "ordinal")
Krippendorff's alpha

Subjects = 12
Raters = 4
alpha = 0.815
> kripp.alpha(a4d12, "interval")
Krippendorff's alpha

Subjects = 12
Raters = 4
alpha = 0.849
    
```

Fig. 5 Computing Krippendorff’s  $\alpha$  using R’s irr library

Table 5 The value-by-unit (i.e., label-by-document) matrix

| Units:  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |                            |
|---------|---|---|---|---|---|---|---|---|---|----|----|----|----------------------------|
| Values: | 1 | 3 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 0  | 2  | 0  | $n_{\bullet 1} = 9$        |
|         | 2 | 0 | 3 | 0 | 0 | 4 | 1 | 0 | 1 | 4  | 0  | 0  | $n_{\bullet 2} = 13$       |
|         | 3 | 0 | 1 | 4 | 4 | 0 | 1 | 0 | 0 | 0  | 0  | 1  | $n_{\bullet 3} = 10$       |
|         | 4 | 0 | 0 | 0 | 0 | 0 | 1 | 4 | 0 | 0  | 0  | 0  | $n_{\bullet 4} = 5$        |
|         | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3  | 0  | 0  | $n_{\bullet 5} = 3$        |
| Totals: | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3  | 2  | 1  | $n_{\bullet \bullet} = 40$ |

Table 6 Coincidence matrices for the data in Table 5: observed (left) and expected (right)

| Values:         | 1               | 2               | 3               | 4               | 5 | $n_{i\bullet}$ |  | Values:         | 1                 | 2                  | 3                  | 4                 | 5                 | $n_{i\bullet}$ |
|-----------------|-----------------|-----------------|-----------------|-----------------|---|----------------|--|-----------------|-------------------|--------------------|--------------------|-------------------|-------------------|----------------|
| 1               | $\frac{7}{4-1}$ | $\frac{4}{4-1}$ | $\frac{1}{4-1}$ | $\frac{1}{4-1}$ | 0 | 9              |  | 1               | $\frac{9*8}{39}$  | $\frac{13*9}{39}$  | $\frac{10*9}{39}$  | $\frac{5*9}{39}$  | $\frac{3*9}{39}$  | 9              |
| 2               | $\frac{4}{4-1}$ | 10              | $\frac{4}{4-1}$ | $\frac{1}{4-1}$ |   | 13             |  | 2               | $\frac{9*13}{39}$ | $\frac{13*12}{39}$ | $\frac{10*13}{39}$ | $\frac{5*13}{39}$ | $\frac{3*13}{39}$ | 13             |
| 3               | $\frac{1}{4-1}$ | $\frac{4}{4-1}$ | 8               | $\frac{1}{4-1}$ | 0 | 10             |  | 3               | $\frac{9*10}{39}$ | $\frac{13*10}{39}$ | $\frac{10*9}{39}$  | $\frac{5*10}{39}$ | $\frac{3*10}{39}$ | 10             |
| 4               | $\frac{1}{4-1}$ | $\frac{1}{4-1}$ | $\frac{1}{4-1}$ | 4               | 0 | 5              |  | 4               | $\frac{9*5}{39}$  | $\frac{13*5}{39}$  | $\frac{10*5}{39}$  | $\frac{5*4}{39}$  | $\frac{3*5}{39}$  | 5              |
| 5               | 0               | 0               | 0               | 0               | 3 | 3              |  | 5               | $\frac{9*3}{39}$  | $\frac{13*3}{39}$  | $\frac{10*3}{39}$  | $\frac{5*3}{39}$  | $\frac{3*2}{39}$  | 3              |
| $n_{\bullet j}$ | 9               | 13              | 10              | 5               | 3 | 40             |  | $n_{\bullet j}$ | 9                 | 13                 | 10                 | 5                 | 3                 | 40             |

### 3.3.3 New Approaches to Conducting Relevance Assessments

Recently, crowdsourcing has become popular as a highly cost-effective means for obtaining relevance assessments (e.g., Alonso et al. 2008; Lease and Yilmaz 2011).

**Table 7** Difference functions for the data in Table 5: *ordinal* $\delta_{ij}^2$  (left) and *interval* $\delta_{ij}^2$  (right)

| Values:         | 1                 | 2                 | 3                 | 4                | 5                 | $n_{i\bullet}$ | Values: | 1              | 2              | 3              | 4              | 5              | $n_{i\bullet}$ |
|-----------------|-------------------|-------------------|-------------------|------------------|-------------------|----------------|---------|----------------|----------------|----------------|----------------|----------------|----------------|
| 1               | 0                 | 11 <sup>2</sup>   | 22.5 <sup>2</sup> | 30 <sup>2</sup>  | 34 <sup>2</sup>   | 9              | 1       | 0              | 1 <sup>2</sup> | 2 <sup>2</sup> | 3 <sup>2</sup> | 4 <sup>2</sup> | 9              |
| 2               | 11 <sup>2</sup>   | 0                 | 11.5 <sup>2</sup> | 19 <sup>2</sup>  | 23 <sup>2</sup>   | 13             | 2       | 1 <sup>2</sup> | 0              | 1 <sup>2</sup> | 2 <sup>2</sup> | 3 <sup>2</sup> | 13             |
| 3               | 22.5 <sup>2</sup> | 11.5 <sup>2</sup> | 0                 | 7.5 <sup>2</sup> | 11.5 <sup>2</sup> | 10             | 3       | 2 <sup>2</sup> | 1 <sup>2</sup> | 0              | 1 <sup>2</sup> | 2 <sup>2</sup> | 10             |
| 4               | 30 <sup>2</sup>   | 19 <sup>2</sup>   | 7.5 <sup>2</sup>  | 0                | 4 <sup>2</sup>    | 5              | 4       | 3 <sup>2</sup> | 2 <sup>2</sup> | 1 <sup>2</sup> | 0              | 1 <sup>2</sup> | 5              |
| 5               | 34 <sup>2</sup>   | 23 <sup>2</sup>   | 11.5 <sup>2</sup> | 4 <sup>2</sup>   | 0                 | 3              | 5       | 4 <sup>2</sup> | 3 <sup>2</sup> | 2 <sup>2</sup> | 1 <sup>2</sup> | 0              | 3              |
| $n_{\bullet j}$ | 9                 | 13                | 10                | 5                | 3                 |                |         |                |                |                |                |                |                |

Furthermore, while Fig. 4 shows the traditional approaches of giving a label to each *document*, some researchers have explored *preference judgements*, where the assessor is asked to enter which of the two documents presented side-by-side is more relevant to the topic (e.g., Carterette et al. 2008a; Chandar and Carterette 2012). Another novel approach is to dynamically present documents and *nuggets* (i.e., pieces of information automatically extracted from the documents) at the same time to the assessor and let him interact with both lists (e.g., Ekstrand-Abueg et al. 2013). This approach aims to identify relevant *information*, not just relevant documents.

## 4 Evaluating Runs

This section describes how ad hoc IR runs can be evaluated based on relevance assessments constructed as described in Sect. 3.3. Section 4.1 describes how evaluation measure scores can be computed. Section 4.2 briefly describes how statistical significance tests can be conducted for comparing different runs. For details on statistical significance testing for IR, the reader is referred to Sakai (2018a). Moreover, while Sect. 4.2 focusses on classical significance tests and *p*-values, Bayesian approaches to hypothesis testing are also available: we refer the reader to Carterette (2015) and Sakai (2017b) for details.

### 4.1 Computing Evaluation Measures

A variety of IR evaluation measures are available: see Sakai (2014) for an overview. In this section, let us evaluate the three runs that we used in Fig. 2 using an existing toolkit, namely, the aforementioned NTCIREVAL.<sup>21</sup> In Sect. 3.2, we used this toolkit just to split the run files into per-topic files; here, we use it to split

<sup>21</sup>See also [http://trec.nist.gov/trec\\_eval/](http://trec.nist.gov/trec_eval/).

```

$ (head -3 wwwE.qrels; tail -3 wwwE.qrels)
0001 clueweb12-0000wb-16-36432 L3
0001 clueweb12-0000wb-31-30391 L1
0001 clueweb12-0000wb-87-27253 L1
0100 clueweb12-1912wb-80-01681 L2
0100 clueweb12-1913wb-11-05574 L2
0100 clueweb12-1913wb-56-22315 L2
$ NTCIRsplitqrels wwwE.qrels rel
created wwwE.qrels.tid
created 0001/0001.rel
created 0002/0002.rel

```

⋮

```

created 0099/0099.rel
created 0100/0100.rel
$ head -3 0100/0100.rel
clueweb12-0000tw-02-05208 L4
clueweb12-0000tw-09-11138 L3
clueweb12-0000wb-27-24543 L4

```

Fig. 6 Creating `rel` files using `NTCIRsplitqrels` from the `NTCIREVAL` toolkit

the relevance assessment file into per-topic files and to compute various evaluation measures for each run.

The file `CLEF20sakai.tar.gz` that we used in Sect. 3.2 contains a sample *query-relevance set* (*qrels*) (Voorhees 2002) file called `wwwE.qrels`, which contains the relevance assessment results for 100 topics. From the top of Fig. 6, it can be observed that this file contains three fields: the first field is the topic ID; the second field is the document ID; and the third field is the *relevance level*.<sup>22</sup> The relevance levels in this particular file contains *L0* (pooled but judged not relevant) through *L4* (highest relevance level). These relevance levels were constructed by having two assessors judge the same topic independently, where each assessor chose from highly relevant (2 points), relevant (1 points), and nonrelevant (0 points). The points were simply summed to form the 5-point relevance scale.

Figure 6 also shows how a script called `NTCIRsplitqrels` may be used to split the `qrels` file into per-topic `rel` files. The second argument to this script (in this case just `rel`) specifies the file suffix for each `rel` file, since there may be several different versions of `rel` (i.e., relevance) files (e.g. `ver1.rel`, `binary.rel`, etc.). The bottom of Fig. 6 shows that a `rel` file just contains document IDs with their relevance levels.

To evaluate the runs, each run file must be split into per-topic `res` files. However, in our case, we have already done that to create the pool files (Sect. 3.2, Fig. 2) and therefore we can proceed directly to evaluation measure calculation. Figure 7 shows how a script called `NTCIR-eval` may be used for this purpose: this script calls a C program called `ntcir_eval`, so make sure you type `make` before using it, as

<sup>22</sup>This is different from the standard TREC format `qrels`, but the two formats are easily interchangeable.

```

$ cat runlist | NTCIR-eval Etidlist rel test -cutoffs 1,10 -g 1:2:3:4
created 0001/0001.RMIT-E-NU-Own-1.test.lab
created 0002/0002.RMIT-E-NU-Own-1.test.lab
.
.
.
created 0099/0099.THUIR-E-PU-Base-3.test.lab
created 0100/0100.THUIR-E-PU-Base-3.test.lab
created THUIR-E-PU-Base-3.test.nev

```

Fig. 7 Creating nev files using NTCIR-eval from the NTCIREVAL toolkit

explained in the README file of NTCIREVAL. The file `runlist` is just a list of the three run file names; the arguments to `NTCIR-eval` are:

First argument List of topic IDs;

Second argument Suffix of the `rel` files to be used;

Third argument A label that you want to add to the names of the output files, i.e., the `nev` (short for `ntcireval`) files and the per-topic `lab` (i.e., labels) files.

In Fig. 7 the label is simply “`test`,” but in practice it is useful to specify a short string that represents a particular experimental condition, such as which version of the `rel` file was used and what gain value setting was used.

Other arguments Arguments directly passed on to `ntcir_eval` the C program.

The `-cutoffs` option specifies one or more measurement depths for depth-based evaluation measures such as *normalized Discounted Cumulative Gain* (*nDCG*): in this example, measures are computed at cutoffs 1 and 10. The `-g` options declares the highest relevance level (*L4* in this example, since four values are specified with this option), and at the same time specifies the *gain value* (Järvelin and Kekäläinen 2002; Sakai 2014) for each relevance level (1 point for *L1*, 2 points for *L2*, etc., in this example).

The per-topic `lab` files are just the `res` files with the relevance levels attached, which are useful for close analysis of each search result: you can see which relevant documents were retrieved at which ranks. The `lab` files may be deleted if not required, but note that looking at the actual ranked lists is more important than looking at numbers (i.e., per-topic evaluation measure scores or the mean over the topic set).

The command shown in Fig. 7 also creates an `nev` file for each run. Figure 8 shows what kind of information the `nev` file for the THUIR run contains. Here, we explain the information provided for Topic 0001 in the `nev` file for THUIR:

`syslen` Size of the ranked list to be evaluated. Measures suffixed with `@l` (e.g., `AP@l`) evaluates the top *l* documents only, while those without `@l` (e.g. `AP`) evaluates the entire `res` file.

`jrel` Number of judged relevant documents (i.e., relevance level *L1* or higher) found in the `res` file.

`jnonrel` Number of judged nonrelevant documents (i.e., *L0* documents) found in the `res` file.

`r1` rank of the first relevant (*L1* or higher) document found in the `res` file.

```

$ head -30 THUIR-E-PU-Base-3.test.nev
0001 # syslen=100 jrel=223 jnonrel=12
0001 # r1=1 rp=1
0001 RR=                1.0000
0001 O-measure=        1.0000
0001 P-measure=        1.0000
0001 P-plus=           1.0000
0001 AP=               0.2410
0001 Q-measure=        0.2257
0001 NCUgu,P=          0.3373
0001 NCUgu,BR=         0.3158
0001 NCurb,P=          0.9280
0001 NCurb,BR=         0.8712
0001 RBP=              0.7998
0001 ERR=              0.8918
0001 AP@0001=          1.0000
0001 Q@0001=           1.0000
0001 nDCG@0001=        1.0000
0001 MSnDCG@0001=      1.0000
0001 P@0001=           1.0000
0001 nERR@0001=        1.0000
0001 Hit@0001=         1.0000
0001 AP@0010=          1.0000
0001 Q@0010=           0.9501
0001 nDCG@0010=        0.9216
0001 MSnDCG@0010=      0.9198
0001 P@0010=           1.0000
0001 nERR@0010=        0.9991
0001 Hit@0010=         1.0000
0002 # syslen=100 jrel=237 jnonrel=3
0002 # r1=1 rp=27

```

Fig. 8 A peek into an nev file

*rp* Preferred rank (Sakai 2014), i.e., the rank of the first *Lh*-relevant document in the *res* file, where *Lh* is the highest relevance level found within that file.

*RR* Reciprocal Rank (*RR*), i.e.,  $1/r_1$  if the *res* file contains at least one relevant document; 0 otherwise.

*O-measure*, *P-measure*, *P-plus* Measures for navigational search intents; extends *RR* for graded relevance (Sakai 2014).

*AP* Average Precision (*AP*) (Buckley and Voorhees 2005).

*Q-measure* A measure similar to average precision; it can handle graded relevance (Sakai 2004).

*NCU* instances of *Normalised Cumulative Utility (NCU)* measures (Sakai and Robertson 2008).

*RBP* Rank-Biased Precision (*RBP*) (Moffat and Zobel 2008).

*ERR* Expected Reciprocal Rank (*ERR*) (Chapelle et al. 2009).

*AP@l*, *Q@l* *AP* and *Q-measure* at cutoff *l* (Sakai 2014).

*nDCG@l* *nDCG* at cutoff *l* as defined in Järvelin and Kekäläinen (2002); not recommended as it does not discount gains for ranks  $r \leq b$ , where *b* is the logarithm base intended as a patience parameter (Sakai 2014).



```

$ Topicsys-matrix Etidlist runlist test.nev MSnDCG@0010 > 100x3.MSnDCG@00
10
$ !!:gs/MSnDCG@/Q@
Topicsys-matrix Etidlist runlist test.nev Q@0010 > 100x3.Q@0010

```

Fig. 9 Creating  $100 \times 3$  topic-by-run matrix files from the three nev files

MSnDCG@/ “Microsoft version” of nDCG as defined in Burges et al. (2005); free from the above problem of the original nDCG, but lacks the patience parameter  $b$ .  
P@/ Precision at cutoff  $l$ .

nERR@/ *normalised Expected Reciprocal Rank (nERR)* at cutoff  $l$ ; normalises ERR based on the *ideal list* (Sakai 2014).

Hit@/ Hit at cutoff  $l$ . That is, 1 iff top  $l$  contains at least one relevant document.

NTCIR-eval can compute other measures, such as those based on a *condensed list* for handling highly incomplete relevance assessments (Sakai 2007), and those designed for *search result diversification* such as *Intent-Aware Expected Reciprocal Rank (ERR-IA)* (Chapelle et al. 2011) and  $D_{\#}$ -nDCG (Sakai and Song 2011).

Figure 9 shows how topic-by-run score matrices can be created from the nev files. In this particular example, a simple script from NTCIREVAL called Topicsys-matrix is utilised to create a Microsoft nDCG and a Q-measure score files. Note that each column represents a run in the order specified in the runlist file: thus, the first column is the RMIT run; the second is the RUCIR run; the third is the THUIR run. We shall utilise these files for statistical significance testing in Sect. 4.2.

## 4.2 Statistical Significance Testing

You have finished evaluating your runs with your favourite evaluation measures and now have topic-by-run score matrices like the ones shown in Fig. 9. Let us take 100x3.MSnDCG@0010 as an example: this file is actually also included in clef20sakai.tar.gz (See Sect. 3.2). By computing the mean for each column, you can see that the Mean nDCG scores for the three runs are 0.6302, 0.5254, 0.5679.

In statistical significance testing, we view the above means as *sample means*, based on a particular (random) sample of topics that we happened to have. Thus, if we have a different sample, we get a different set of sample means. What, then, are the *true* means, or the *population means* where we consider all possible topics? Are the population means really different?

```

$ Random-test runlist 100x3.MSnDCG@0010 10000
created 100x3.MSnDCG@0010.pvalues.10000
$ cat 100x3.MSnDCG@0010.pvalues.10000
RMIT-E-NU-Own-1 RUCIR-E-NU-Base-1 0.104783 0.0006
RMIT-E-NU-Own-1 THUIR-E-PU-Base-3 0.062255 0.045
RUCIR-E-NU-Base-1 THUIR-E-PU-Base-3 0.042528 0.2444
$
$ Random-test runlist 100x3.MSnDCG@0010 50000
created 100x3.MSnDCG@0010.pvalues.50000
$ cat 100x3.MSnDCG@0010.pvalues.50000
RMIT-E-NU-Own-1 RUCIR-E-NU-Base-1 0.104783 0.00022
RMIT-E-NU-Own-1 THUIR-E-PU-Base-3 0.062255 0.046
RUCIR-E-NU-Base-1 THUIR-E-PU-Base-3 0.042528 0.24246

```

**Fig. 10** Conducting an RTHSD test with  $B = 10,000$  and  $B = 20,000$  trials

If you are only interested in the difference between a particular *pair* of systems (e.g., RMIT vs RUCIR), a paired  $t$ -test can be applied.<sup>23</sup> This is very easy to do using R or even Microsoft Excel (Sakai 2018a). If, on the other hand, you are interested in the difference between *every system pair*, applying the  $t$ -test independently for every system pair is *not* the correct approach, as this would inflate the *familywise Type I Error rate*, i.e., the probability that at least one of the significance test detects a difference that is not real (Carterette 2012; Sakai 2018a). The correct approach is to use a *multiple comparison procedure*.

The popular *Bonferroni correction*, which divides the significance criterion  $\alpha$  by the number of independent significance tests to prevent the inflation of the familywise error rate, should now be considered obsolete (Crawley 2015). An example of better multiple comparison procedures would be *Tukey’s Honestly Significant Difference test*, which ensures that the familywise error rate is bounded above by  $\alpha$  (say, 0.05). This test is also available in R (function `TukeyHSD`) (Sakai 2018a). It should also be noted that there is no need to conduct ANOVA prior to conducting the Tukey HSD test; See Sakai (2018a) Chapter 4 for a discussion.

Here, let us discuss the *Randomised Tukey Honestly Significant Difference (RTHSD)* test (Carterette 2012; Sakai 2018a), which is a randomisation test version of the aforementioned Tukey test. Unlike classical significance tests, RTHSD does not rely on the random sampling assumption: its null hypothesis merely assumes that the scores of all systems actually came from the same “hidden” system. Based on this assumption, it generates a null distribution by permutating the rows of the topic-by-run matrix. See Sakai (2018a) for more details; below, we only discuss how to conduct the RTHSD test.

To conduct an RTHSD test with a topic-by-run matrix, you can download the `Discpower` toolkit.<sup>24</sup> Details can be found in README. Figure 10 shows how the `Random-test` script can be used with the aforementioned topic-by-run file

<sup>23</sup>Computer-based, distribution-free alternatives to the  $t$ -test can also be applied: the *bootstrap test* (Sakai 2006) and the *randomisation test* (Smucker et al. 2007).

<sup>24</sup><http://research.nii.ac.jp/ntcir/tools/discpower-en.html>.

100x3.MSnDCG@0010: the first argument specifies the run names that correspond to the three columns of the topic-by-run file; the third argument is the number of trials (i.e., how many permuted matrices are created) for computing the  $p$ -value.

Because RTHSD exploits computer power instead of relying on a mathematical definition for obtaining the null distribution, executing `Random-test` takes a while (as in, several hours or even more, depending on the matrix size and the number of trials  $B$ ). Figure 10 tries  $B = 10,000$  and  $B = 50,000$ , although in practice you just need to choose a value for  $B$ . The output files of `Random-test` contain the run pairs, the absolute difference, and finally the  $p$ -value; it can be observed that the  $p$ -values are slightly different depending on  $B$ , although the difference will not affect the dichotomous statistical significance decisions at (say)  $\alpha = 0.05$ . To obtain sufficiently accurate  $p$ -values, I recommend at least 5000 trials.

The  $p$ -value is the probability of observing the difference in sample means that you have observed *under the assumption that the two population means are actually equal*. The use of the  $p$ -value has its limitations: it is a function not only of the *effect size* (i.e., the magnitude of the difference between two systems) but also of the *sample size* (i.e., how many topics we have). That is, it is possible to obtain arbitrarily low  $p$ -values by increasing the sample size. Hence, reporting the effect size along with the  $p$ -value is encouraged. Thus, for each pair of runs  $(i, i')$ , an effect size given as a *standardised mean difference* can be computed as:

$$ES_{E2}(i, i') = \frac{\bar{x}_{i\bullet} - \bar{x}_{i'\bullet}}{\sqrt{V_{E2}}} \quad (16)$$

where  $\bar{x}_{i\bullet}$ ,  $\bar{x}_{i'\bullet}$  the sample means for runs  $i$ ,  $i'$ , respectively, and  $V_{E2}$  is obtained from the topic-by-run matrix using Eq. (1). That is, Eq. (16) measures the difference in standard deviation units. See Sakai (2018a) for more details.

## 5 Impacting the Research Community

This is the final section of this chapter, which discusses what should happen after completing the evaluation of your task. Section 5.1 discusses publications and data release based on your task; Sect. 5.2 discusses redesigning the task based on your experience; finally, Sect. 5.3 discusses the impact of evaluation tasks on people.

### 5.1 Publishing Papers, Releasing the Data

You should write a high-quality, detailed overview paper for your task. Start with your motivation and clear task definition as we have discussed in Sect. 1. If you wrote a good task proposal document, then you can reuse a lot of material from that. Provide the detailed evaluation results, complete with statistical significance

testing and effect size results. More importantly, conduct per-topic failure analysis and visualise the results. You should also encourage your participants to write high-quality papers that provide the details of their runs and their effectiveness.

Paper publishing should go beyond the evaluation venues such as CLEF, NTCIR, and TREC: submit papers to top conferences and journals, so that more researchers will be interested in your work and may join the future rounds.

Release the task data as much as possible. Instead of just publishing the mean nDCG results in your overview paper, release the topic-by-run matrices, or even better, the raw run files, to facilitate replications of, and improvements on, the experiments done in your task. Making your data public is extremely important for making your task impactful in the research community.

## 5.2 Improving the Task and Monitoring Progress

You must have learnt something from your task. Exploit it to improve the task design. However, drastically changing the task design for the next round is risky, as this may set a high bar on participants who want to come back to the task. Having a new pilot subtask along with a more conventional main task is one safe approach.

If you created your test collection based on topic set size design (See Sect. 3.1), you can consider a new test collection design for the next round of your task. You have a new topic-by-run matrix for your favourite evaluation measure as a fruit of running the task; so you can obtain a new variance estimate from it using Eq. (1). If the new matrix is larger than the pilot data you used initially, the new variance estimate is more trustworthy. If the pilot and the new matrices are similar in scale, you may obtain a *pooled variance* for a new topic set size design. If the number of topics for a topic-by-run matrix obtained from a collection  $C$  is denoted by  $n_C$  and the variance estimate based on it (obtained using Eq. (1)) is denoted by  $\hat{\sigma}_C^2$ , the pooled variance can be obtained as (Sakai 2016):

$$\hat{\sigma}^2 = \frac{\sum_C (n_C - 1) \hat{\sigma}_C^2}{\sum_C (n_C - 1)}. \quad (17)$$

Ask your participants to “freeze” their systems so that they can be used to process the new topics in the next round of your task. Such runs are sometimes called *revived runs* (Sakai et al. 2013). If a research group comes back to the next round of your task and submit revived runs as well as runs based on their new approaches, then their progress across the two rounds can be quantified on the new topic set. Moreover, this will increase the number of runs for your pools.

### 5.3 *Power to the People*

Evaluation tasks can give (nonstatistical) power to the people. Speaking of my own personal experience, I have participated in various NTCIR tasks since NTCIR-1, which took place in 1999. I made many new friends at the NTCIR conferences, and started running my own tasks by collaborating with others in around 2007. Through my experience as a task participant and as a task organiser, I started designing my own evaluation measures, and ways to evaluate evaluation measures. This led me to work with researchers from the TREC community to run a TREC track in 2013 and 2014. For NTCIR, I have also served as a programme co-chair as well as a general co-chair. Recently, I have started working with Nicola Ferro and Ian Soboroff on a “metatask” that spans CLEF, NTCIR, and TREC.<sup>25</sup> And I am highly confident that I am not the only researcher who grew through IR evaluation tasks.

Evaluation tasks can bring together people from diverse backgrounds. As a result, interdisciplinary research topics, and hence novel evaluation tasks, can be born. Evaluation tasks are more about collaboration than competition. Running them properly requires a lot of effort, but you will be rewarded not only with interesting research findings and publications, but also with a new community in which you are surrounded by friends, old and new.

## References

- Allan J, Carterette B, Aslam JA, Pavlu V, Dachev B, Kanoulas E (2008) Million query track 2007 overview. In: Proceedings of TREC 2007
- Alonso O, Rose DE, Stewart B (2008) Crowdsourcing for relevance evaluation. *SIGIR Forum* 42(2):9
- Bailey P, Craswell N, Soboroff I, Thomas P, de Vries AP, Yilmaz E (2008) Relevance assessment: are judges exchangeable and does it matter? In: Proceedings of ACM SIGIR 2008, pp 667–674
- Buckley C, Voorhees EM (2004) Retrieval evaluation with incomplete information. In: Proceedings of ACM SIGIR 2004, pp 25–32
- Buckley C, Voorhees EM (2005) Retrieval system evaluation. In: Voorhees EM, Harman DK (eds) *TREC: experiment and evaluation in information retrieval*. The MIT Press, Boston, chap 3
- Burges C, Shaked T, Renshaw E, Lazier A, Deeds M, Hamilton N, Hullender G (2005) Learning to rank using gradient descent. In: Proceedings of ACM ICML 2005, pp 89–96
- Carterette B (2012) Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM TOIS* 30(1):4
- Carterette B (2015) Bayesian inference for information retrieval evaluation. In: Proceedings of ACM ICTIR 2015, pp 31–40
- Carterette B, Bennett PN, Chickering DM, Dumais ST (2008a) Here or there: preference judgments for relevance. In: Proceedings of ECIR 2008 (LNCS), vol 4956, pp 16–27
- Carterette B, Pavlu V, Kanoulas E, Aslam JA, Allan J (2008b) Evaluation over thousands of queries. In: Proceedings of ACM SIGIR 2008, pp 651–658

---

<sup>25</sup> CLEF/NTCIR/TREC REproducibility: <http://www.centre-eval.org/>.

- Chandar P, Carterette B (2012) Using preference judgments for novel document retrieval. In: Proceedings of ACM SIGIR 2012, pp 861–870
- Chapelle O, Metzler D, Zhang Y, Grinspan P (2009) Expected reciprocal rank for graded relevance. In: Proceedings of ACM CIKM 2009, pp 621–630
- Chapelle O, Ji S, Liao C, Velipasoglu E, Lai L, Wu SL (2011) Intent-based diversification of web search results: metrics and algorithms. *Inf Retr* 14(6):572–592
- Cohen J (1960) A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20:37–46
- Cohen J (1968) Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol Bull* 70(4):213–220
- Crawley MJ (2015) *Statistics: an introduction using R*, 2nd edn. Wiley, Chichester
- Ekstrand-Abueg M, Pavlu V, Kato MP, Sakai T, Yamamoto T, Iwata M (2013) Exploring semi-automatic nugget extraction for Japanese one click access evaluation. In: Proceedings of ACM SIGIR 2013, pp 749–752
- Fleiss JL (1971) Measuring nominal scale agreement among many raters. *Psychol Bull* 76(5):378–382
- Harman DK (2005) The TREC test collections. In: Voorhees EM, Harman DK (eds) *TREC: experiment and evaluation in information retrieval*. The MIT Press, Boston, chap 2
- Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of IR techniques. *ACM TOIS* 20(4):422–446
- Krippendorff K (2013) *Content analysis: an introduction to its methodology*, 3rd edn. SAGE Publications, Los Angeles
- Lease M, Yilmaz E (2011) Crowdsourcing for information retrieval. *SIGIR Forum* 45(2):66–75
- Luo C, Sakai T, Liu Y, Dou Z, Xiong C, Xu J (2017) Overview of the NTCIR-13 we want web task. In: Proceedings of NTCIR-13
- Moffat A, Zobel J (2008) Rank-biased precision for measurement of retrieval effectiveness. *ACM TOIS* 27(1):2
- Nagata Y (2003) How to design the sample size (in Japanese). Asakura Shoten
- Randolph JJ (2005) Free-marginal multirater kappa (multirater  $\kappa_{free}$ ): an alternative to Fleiss' fixed marginal multirater kappa. In: Joensuu learning and instruction symposium 2005
- Sakai T (2004) Ranking the NTCIR systems based on multigrade relevance. In: Proceedings of AIRS 2004 (LNCS), vol 3411, pp 251–262
- Sakai T (2006) Evaluating evaluation metrics based on the bootstrap. In: Proceedings of ACM SIGIR 2006, pp 525–532
- Sakai T (2007) Alternatives to bpref. In: Proceedings of ACM SIGIR 2007, pp 71–78
- Sakai T (2014) Metrics, statistics, tests. In: PROMISE winter school 2013: bridging between information retrieval and databases (LNCS), vol 8173, pp 116–163
- Sakai T (2015) *Information access evaluation methodology: for the progress of search engines (in Japanese)*. Corona Publishing, New York
- Sakai T (2016) Topic set size design. *Inf Retr J* 19(3):256–283
- Sakai T (2017a) The effect of inter-assessor disagreement on IR system evaluation: a case study with lancers and students. In: Proceedings of EVIA 2017, pp 31–38
- Sakai T (2017b) The probability that your hypothesis is correct, credible intervals, and effect sizes for ir evaluation. In: Proceedings of ACM SIGIR 2017, pp 25–34
- Sakai T (2018a) *Laboratory experiments in information retrieval: sample sizes, effect sizes, and statistical power*. Springer, Cham. <https://link.springer.com/book/10.1007/978-981-13-1199-4>
- Sakai T (2018b) Topic set size design for paired and unpaired data. In: Proceedings of ACM ICTIR 2018
- Sakai T, Lin CY (2010) Ranking retrieval systems without relevance assessments: revisited. In: Proceedings of EVIA 2010, pp 25–33
- Sakai T, Robertson S (2008) Modelling a user population for designing information retrieval metrics. In: Proceedings of EVIA 2008, pp 30–41
- Sakai T, Song R (2011) Evaluating diversified search results using per-intent graded relevance. In: Proceedings of ACM SIGIR 2011, pp 1043–1052

- Sakai T, Dou Z, Yamamoto T, Liu Y, Zhang M, Song R, Kato MP, Iwata M (2013) Overview of the NTCIR-10 INTENT-2 task. In: Proceedings of NTCIR-10, pp 94–123
- Smucker MD, Allan J, Carterette B (2007) A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of ACM CIKM 2007, pp 623–632
- Sparck Jones K, van Rijsbergen CJ (1975) Report on the need for and provision of an ‘ideal’ information retrieval test collection. Tech. rep., Computer Laboratory, University of Cambridge, British Library Research and Development Report No. 5266
- Voorhees EM (2000) Variations in relevance judgments and the measurement of retrieval effectiveness. *Inf Process Manag* 36:697–716
- Voorhees EM (2002) The philosophy of information retrieval evaluation. In: Proceedings of ECIR 2002 (LNCS), vol 2406, pp 355–370
- Webber W, Moffat A, Zobel J (2008) Statistical power in retrieval experimentation. In: Proceedings of ACM CIKM 2008, pp 571–580
- Yilmaz E, Aslam JA (2006) Estimating average precision with incomplete and imperfect judgments. In: Proceedings of ACM CIKM 2006, pp 102–111
- Zobel J (1998) How reliable are the results of large-scale information retrieval experiments? In: Proceedings of ACM SIGIR 1998, pp 307–314

**Part II**  
**Evaluation Infrastructures**



# An Innovative Approach to Data Management and Curation of Experimental Data Generated Through IR Test Collections



Maristella Agosti, Giorgio Maria Di Nunzio, Nicola Ferro,  
and Gianmaria Silvello

**Abstract** This paper describes the steps that led to the invention, design and development of the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* system for managing and accessing the data used and produced within experimental evaluation in *Information Retrieval (IR)*. We present the context in which DIRECT was conceived, its conceptual model and its extension to make the data available on the Web as *Linked Open Data (LOD)* by enabling and enhancing their enrichment, discoverability and re-use. Finally, we discuss possible further evolutions of the system.

## 1 Introduction

Experimental evaluation is a fundamental topic of *Information Retrieval (IR)* and it has the Cranfield paradigm (Cleverdon 1997) at its core. The two key components of experimental evaluation are experimental collections and evaluation campaigns organized at an international level. The management of experimental collections—i.e. documents, topics and relevance judgments—and of the data produced by the evaluation campaigns—i.e. runs, measures, descriptive statistics, papers and reports—are of central importance to guarantee the possibility of conducting evaluation experiments that are repeatable and that permit re-usability of the collections.

A crucial aspect for IR evaluation is to ensure the best exploitation and interpretation, over large time spans, of the used and produced experimental data. Nevertheless, this aspect has often been overlooked in the field, since researchers are generally more interested in developing new algorithms and methods rather than modeling and managing the experimental data (Agosti et al. 2007b,c).

---

M. Agosti · G. M. Di Nunzio (✉) · N. Ferro · G. Silvello  
Department of Information Engineering, University of Padua, Padua, Italy  
e-mail: [maristella.agosti@unipd.it](mailto:maristella.agosti@unipd.it); [giorgiomaria.dinunzio@unipd.it](mailto:giorgiomaria.dinunzio@unipd.it); [nicola.ferro@unipd.it](mailto:nicola.ferro@unipd.it);  
[gianmaria.silvello@unipd.it](mailto:gianmaria.silvello@unipd.it)

As a consequence, within the *Conference and Labs of the Evaluation Forum (CLEF)* evaluation campaigns, we worked on modeling the IR experimental data and on designing a research infrastructure able to manage, curate and grant access to them. This effort led to the invention, design and development of the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* system (Agosti et al. 2012; Ferro et al. 2011) and it raised awareness of the importance of curating and managing research data in the community and beyond (Agosti et al. 2009, 2013, 2014; Allan et al. 2012; Zobel et al. 2011).

DIRECT enables the typical IR evaluation workflow, and manages the scientific data used and produced during large-scale evaluation campaigns. In addition, DIRECT has the potential to support the archiving, access, citation, dissemination and sharing of the experimental results.

On the top of DIRECT, we successively added some *Linked Open Data (LOD)* functionalities (Heath and Bizer 2011)—i.e. the LOD-DIRECT system—to enable the discoverability, enrichment and the interpretability of the experimental data. We defined a *Resource Description Framework (RDF)* model of the IR scientific data also modelling their connections with the scientific papers related and based on them. We also provided a methodology for automatically enriching the data by exploiting relevant external entities from the LOD cloud (Silvello et al. 2017).

The paper is organized as follows: Sect. 2 introduces the complex and rich field of experimental evaluation and the Cranfield paradigm, and provides the scientific context in which DIRECT has been invented, designed and developed. Section 3 presents the conceptual model of the infrastructure, and the main conceptual areas composing it, highlighting how experimental data are modeled within the system. Section 4 describes the semantic model defined for publishing IR experimental data on the Web as LOD and LOD-DIRECT. Section 5 refers to related work. Finally, Sect. 6 discusses and considers possible future developments.

## 2 The Cranfield Paradigm and the Evaluation Campaigns

### 2.1 Abstraction of IR Systems Evaluation

The evaluation of information retrieval systems is an abstraction of the retrieval process based on a set of choices that represent certain aspects of the real world (directly or indirectly) and ignore others (Robertson 2008). This abstraction allows researchers in IR to control some of the variables that affect retrieval performance and exclude other variables that may affect the noise of laboratory evaluation (Voorhees 2002). The “Cranfield paradigm” is at the heart of the design of laboratory experiments of evaluation of information retrieval tools and systems (Cleverdon 1997; Harman 2011). This paradigm defines the notion of the methodology of experimentation in IR, where the goal is to create “a laboratory type situation where, freed as far as possible from the contamination of

operational variables, the performance of index languages could be considered in isolation” (Cleverdon 1997). The core of this methodology abstracts away from the details of particular tasks and users and instead focuses on a benchmark called “test collection” which consists of three components: a set of documents, a set of topics, and a ground truth, i.e. a set of relevance assessments for each document-topic pair. The abstracted retrieval task is to rank the document set for each topic, then the effectiveness of a system for a single topic is computed as a function of the ranks of the relevant documents (Voorhees 2007).

Some years after the Cranfield paradigm was established, researchers in the field of IR noted that the collections existing at that time, which had been designed and created for a specific experimental evaluation of a system and/or a comparison between systems, were re-used for many other experiments, for which they were not ideal (Spärck Jones and van Rijsbergen 1975; Spärck Jones and Bates 1977). Some of the issues were related to the lack of suitable test data and the way that the experiments were documented often without suitable caveats. Quoting a passage from Spärck Jones and van Rijsbergen (1975):

There is a widespread feeling among research workers that existing test collections are inadequate because they are small and/or careless and/or inappropriate. They may also not be fully machine-readable, or may be in an esoteric machine format.

On the basis of these considerations, Karen Spärck Jones and Keith van Rijsbergen clarified and illustrated the characteristics that an ‘ideal’ test collection must have to overcome the aforementioned problems (Spärck Jones and van Rijsbergen 1975; Robertson 2008).

## 2.2 *The Ideal Test Collection and TREC*

The concept of an ideal test collection was implemented for the first time in the context of the first *Text REtrieval Conference (TREC)*<sup>1</sup> in 1992, that is many years after its definition. One of the goals of TREC has been to provide a shared task evaluation that allows cross-system comparisons. In addition to the initial traditional ad-hoc task, there have been a wide range of experimental tracks that have focused on new areas or particular aspects of text retrieval since TREC 4 (Harman 1995). To adhere to the ideal test collection characteristics, for each TREC track participants receive: a collection of documents obtained from some external source; a collection of topics, which may also be obtained externally or may be created internally; and a set of relevance assessments, known as *qrels*. Each participant tests a search system on the collection and produces as a result a ranked list of documents for each topic—known as a *run*—which is submitted to NIST. The runs submitted by the participants are pooled in order to produce the set of relevance assessment (Robertson 2008).

---

<sup>1</sup><http://trec.nist.gov/>.

Since its beginning in 1992, the TREC effort has had a profound influence on all aspects of evaluation, from the formatting of test collection documents, topics, and grels, through the types of information needs and relevance judgments made, to the precise definition of evaluation measures used (Voorhees and Harman 2005; Sanderson 2010). In addition to experimental collection material, TREC has also greatly encouraged the development of good methods of experimentation. The standard of rigour of experimental methodology has been vastly improved (Robertson 2008) thanks to TREC, which has become a yearly evaluation initiative, or evaluation campaign, of reference for all academic and industrial communities of information retrieval.

### 2.3 *The Management of Data Produced in the Context of Evaluation Campaigns*

Donna Harman and her colleagues appeared to be the first to realize that if the documents and topics of a test collection were distributed for little or no cost, a large number of groups would be willing to use that data in their search systems and submit runs back to TREC at no cost (Sanderson 2010). Moreover, the materials and methods TREC has generated are materials and methods for laboratory experiments (Robertson 2008). In this respect, since its beginning TREC has promoted the concept of reusability which facilitates research.

Despite the fact that IR has traditionally been very rigorous about experimental evaluation, researchers in this field have raised some concerns about the reproducibility of system experiments because, among other things, there is not a clear methodology for managing experimental data across different conferences and evaluation initiatives (Ferro 2017). In fact, after TREC, other evaluation campaigns were launched to deal with the evaluation of many different IR approaches and systems that were being defined also thanks to the development of many different types of IR systems and tools, such as, for example, Web search engines.

Some important relevant evaluation campaigns that have been launched over the years and that are still active now are: NTCIR (NII Testbeds and Community for Information access Research), Japan, from 1999<sup>2</sup>; CLEF (Conference and Labs of the Evaluation Forum), Europe, from 2000<sup>3</sup>; FIRE (Forum for Information Retrieval Evaluation), India, from 2008.<sup>4</sup>

The *INitiative for the Evaluation of XML Retrieval (INEX)* has provided the means to evaluate focused retrieval search engines, especially *eXtensible Markup*

---

<sup>2</sup><http://research.nii.ac.jp/ntcir/index-en.html>.

<sup>3</sup><http://www.clef-initiative.eu/>.

<sup>4</sup><http://fire.irsi.res.in/>.

*Language (XML)* retrieval; it was launched in 2002, came under the CLEF umbrella in 2012, but ran for the last time in 2014.<sup>5</sup>

As reported, many evaluation initiatives are active and produce important results for the evaluation of IR systems and tools. Naturally, these different initiatives have been launched and conducted to respond to different research questions, so they have specificities that do not always make cross-comparability between the initiatives possible. As a consequence, the produced experimental results are often not cross comparable. We started from this consideration to work on proposing a conceptual model of an infrastructure that would face and solve some of the problems related to the management and curation of the data produced during an evaluation campaign (Agosti et al. 2007a,c).

### 3 Conceptual Model of the Infrastructure

In IR, as well as in other related scientific fields, a crucial topic that has to be addressed is how to guarantee that the data produced by the scientific activities are consistently managed, are made accessible and available for re-use and are documented to make them easily interpretable. In IR evaluations, these are key aspects, and especially in the context of large evaluation campaigns such as CLEF. For example, the importance of describing and annotating scientific datasets is discussed in Bowers (2012), noting that this is an essential step for the interpretation, sharing, and reuse of the datasets.

We thus began an exercise aimed at modeling the IR experimental data and designing a software infrastructure able to manage and curate them, which led to the development of the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* system (Di Nunzio and Ferro 2005; Agosti et al. 2012). This effort contributed to raising awareness and consensus in the research community and beyond (Agosti et al. 2009; Allan et al. 2012; Forner et al. 2013; Zobel et al. 2011; Ferro et al. 2011).

DIRECT models all the aspects of a typical evaluation workflow in IR and provides the means to deal with some advanced aspects that have been receiving attention in recent years, such as bibliometrics based on data and the visualization of scientific data.

We can model the main phases of the IR experimental evaluation workflow as follows:

- The first phase regards the creation of the experimental collection composed of the acquisition and preparation of the documents (*D*) and the creation of topics (*T*) from which a set of queries is generated.

---

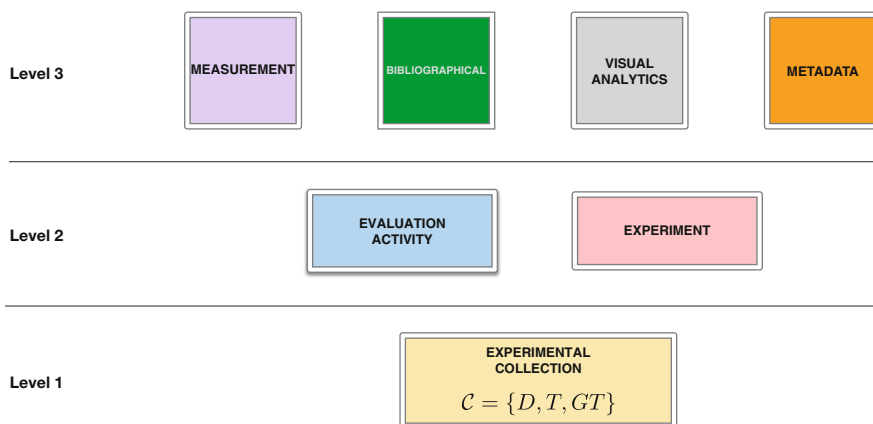
<sup>5</sup><https://inex.mmci.uni-saarland.de/>.

- The second phase concerns the participants in the evaluation campaign who run experiments and test their systems.
- In the third phase, the experiments are gathered and used by the campaign organizers to create the ground-truth ( $GT$ ).
- In the fourth phase, measurements are calculated.
- In the fifth phase the measurements are used to produce descriptive statistics and conduct statistical tests about the behavior of one or more systems.
- The sixth and last phase regards the scientific production where both participants and organizers prepare reports about the campaign and the experiments, the techniques they used, and their findings. This phase usually continues also after the conclusion of the campaign as the investigations of the experimental results require a deeper understanding and further analyses which may lead to the production of conference and journal papers.

The conceptual schema of the infrastructure, abstracting from the actual phases of the IR experimental evaluation workflow, models the evaluation workflow by means of seven functional areas organized in three main conceptual levels; Fig. 1 provides an intuitive representation of them. The three levels are built one on top of the other since the experimental collection area constitutes the basis of the evaluation activities and the experiments on level 2. In the same fashion, the measurement, bibliographical, visual analytics and metadata areas on level 3 depend on the areas on level 2.

We document in the following the aim and the content of each functional area.

**Experimental Collection Area** This area belongs to the first conceptual level and it allows us to set up a traditional IR evaluation environment following the classic Cranfield paradigm based on the triple  $\mathcal{C} = \{D, T, GT\}$ : a corpus of documents, a group of topics and a set of assessments on the documents with regard to the considered topics. In the abstraction process particular attention has been paid to



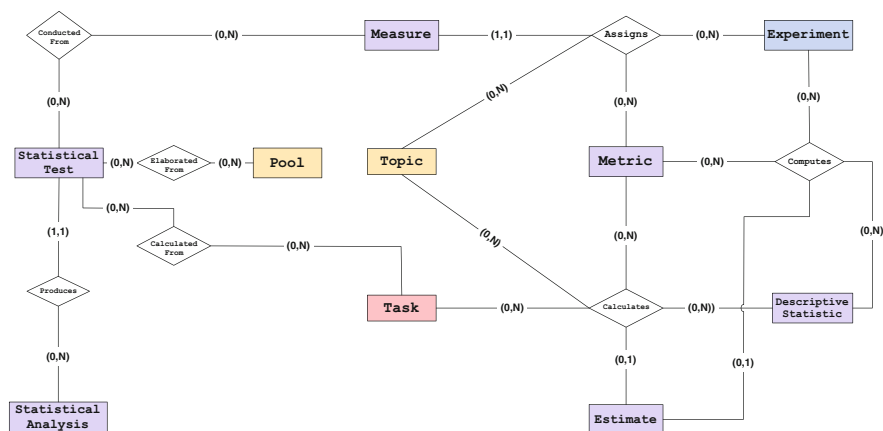
**Fig. 1** The conceptual areas of the evaluation infrastructure

the concept of *topic*, because of the diversity of the information needs that have to be addressed in different evaluation tasks.

**Evaluation Activity Area** This area belongs to the second conceptual level and builds on the experimental collection area. It identifies the core of the infrastructure; it refers to activities aimed at evaluating applications, systems and methodologies for multimodal and multimedia information access and retrieval. Entities in this area go beyond the traditional evaluation campaigns by including trial and education activities. *Trial* refers to an evaluation activity that may be actively run by, say, a research group, a person or a corporate body for their own interest. This evaluation activity may or may not be shared with the community of interest; for instance, a trial activity may be the experiments performed to answer a research question and to write a research paper or the activities conducted to evaluate a Web application. The *Education* activities allow us to envision evaluation activities carried out for educational purposes. In a certain sense, this area extends the activities considered by the Cranfield paradigm.

**Experiment Area** This area belongs to the second conceptual level and concerns the scientific data produced by an experiment carried out during an evaluation activity. Also in this case, this area models the traditional Cranfield experimental settings and extends it by allowing other side evaluation activities. Indeed, the evaluation infrastructure considers three different types of experiment: run, guerrilla, and living. A *Run*, produced by an IR system, is defined as a ranked list of documents for each topic in the experimental collection (Voorhees and Harman 2005) in a classic IR evaluation context. A *Guerrilla* experiment identifies an evaluation activity performed on corporate IR systems (e.g. a custom search engine integrated in a corporate Web site) (Agosti et al. 2012); in a guerrilla experiment, the evaluation process is defined by a set of experimental activities aimed at assessing different aspects of the application, such as the completeness of the index of an ad-hoc search engine or the effectiveness of the multilingual support. For this reason the evaluation metrics may differ from those used during a Run experiment. A *Living* experiment deals with the specific experimental data resulting from the Living Retrieval Laboratories, which examines the use of operational systems on an experimental platform on which to conduct user-based experiments to scale.

**Measurement Area** This area belongs to the third conceptual level and concerns the measures used for evaluation activities. This area is one of the most important of the infrastructure and it constitutes one element of distinction between DIRECT and other modeling efforts in the IR evaluation panorama. In Fig. 2 we can see relationships among the main entities of this area and other entities in the evaluation activity, the experimental collection, and the experiment area. For a topic-experiment pair a specific value of a metric, namely a measure, is assigned—i.e. a Measure refers to one and only one Experiment-Topic-Metric triple through the relationship Assigns. If we consider the results on an experiment basis, then Descriptive Statistics can be computed for a given Metric. Descriptive Statistics can be computed also on a task



**Fig. 2** The ER schema modeling the measurement area

basis. A Statistical Analysis can produce a value for a specific statistical test; the Statistical Test value can be Elaborated From data in none, one or more Pools, or Calculated From data from none, one or more Tasks, or Computed From an Experiment.

The main point here is that explicitly considering the entities in the measurement area as a part of the conceptual schema we are able to retain and make accessible not only experimental data, but also evaluation methodologies and the context wherein metrics and methodologies have been applied.

**Metadata Area** This area belongs to the third level and supports the description and the enrichment through metadata of the resources handled by the infrastructure. Generally, metadata describing the resources are not considered central resources in an evaluation infrastructure: whereas, in DIRECT they are considered important resources and managed alongside the other classical evaluation resources. This allows us to use metadata in concert with measures and experiments for enriching the experimental data as we discuss below.

**Bibliographical Area** This area belongs to the third level and it is responsible for making explicit and retaining the relationship between the data that result from the evaluation activities and the scientific production based on these data. This area is central for dealing with bibliometrics of experimental data and for dealing with data provenance (Buneman et al. 2000) and citation (Davidson et al. 2017).

**Visual Analytics Area** This area belongs to the third level and it manages the information used by the infrastructure to store and recover whatever visualization of the data that users produce. This area manages the information used by the infrastructure to store and retrieve parametric and interactive visualizations of the data.



To the best of our knowledge, DIRECT is the most comprehensive tool for managing all the aspects of the IR evaluation methodology, the experimental data produced and the connected scientific contributions. Besides supporting the design of an innovative evaluation infrastructure, another goal of DIRECT is to provide a common abstraction of IR evaluation activities that can be exploited to share and re-use the valuable scientific data produced by experiments and analysis and to envision evaluation activities other than traditional IR campaigns.

## 4 A Semantic Mapping of the Conceptual Model

Research data are of key importance across all scientific fields as these data constitute a fundamental building block of the system of science. Recently, a great deal of attention has been dedicated to the nature of research data and how to describe, share, cite and re-use them in order to enable reproducibility in science and to ease the creation of advanced services based on them (Borgman 2015; Silvello 2017). In this context, the *Linked Open Data (LOD)* paradigm (Heath and Bizer 2011) is a de-facto standard for publishing and enriching data; it allows the opening-up of public data in machine-readable formats ready for consumption, re-use and enrichment through semantic connections enabling new knowledge creation and discovery possibilities. The LOD paradigm can be mainly seen as a method of publishing structured data so that data can be interlinked. It builds upon standard Web technologies such as *HyperText Transfer Protocol (HTTP)* and *RDF*,<sup>6</sup> but rather than using them to serve web pages for humans, “it extends them to share information in a way that can be read automatically by machines”.<sup>7</sup>

In the field of IR, the LOD paradigm is not as central as it is in other fields such as life science research (Gray et al. 2014) and social sciences (Zapilko et al. 2013). So, despite the centrality of data, in IR there are no shared and clear ways to publish, enrich and re-use experimental data as LOD with the research community.

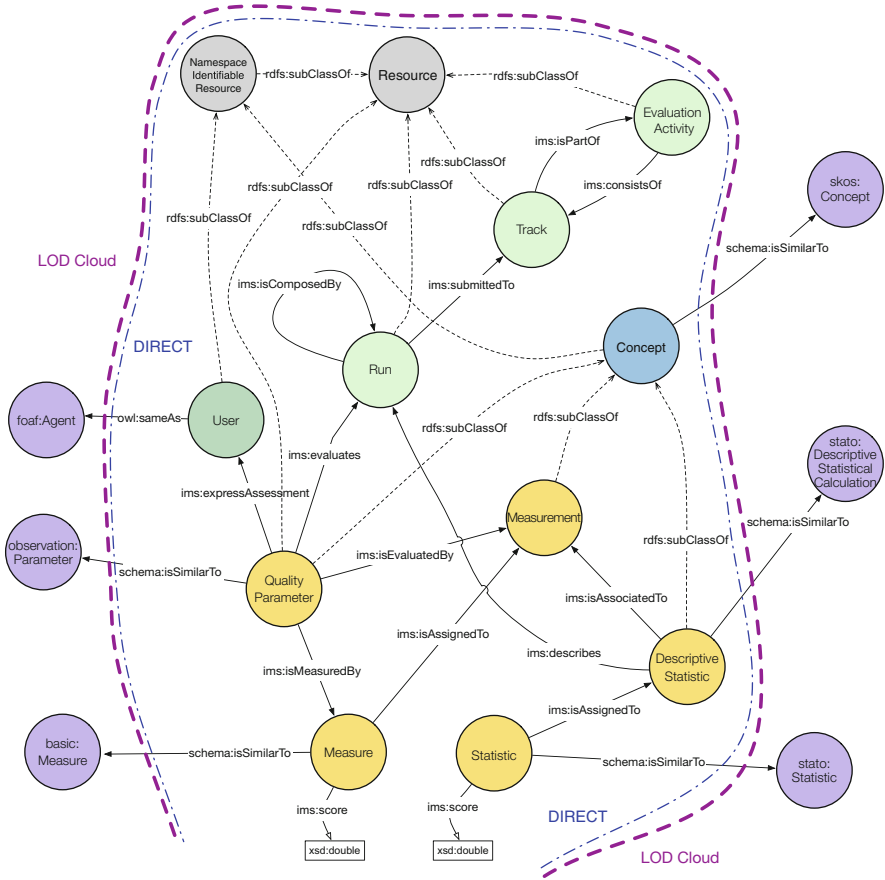
To target this aspect of data sharing, re-use and enrichment within the DIRECT infrastructure, we defined an RDF model (W3C 2004) for representing experimental data and publishing them as LOD on the Web. This can enable seamless integration of datasets produced by different experimental evaluation initiatives as well as the standardization of terms and concepts used to label data across research groups and interested organizations (Silvello et al. 2017).

Moreover, with the purpose of augmenting the access points to the data as well as the potential for their interpretability and re-usability, we built upon the proposed RDF model to automatically find topics in the scientific literature, exploiting the scientific IR data as well as connecting the dataset with other datasets in the LOD cloud.

---

<sup>6</sup><https://www.w3.org/standards/semanticweb/data>.

<sup>7</sup>[https://en.wikipedia.org/wiki/Linked\\_data](https://en.wikipedia.org/wiki/Linked_data).



**Fig. 3** The experiment area classes and properties

The detection of scientific topics related to the data produced by the experimental evaluation and the enrichment of scientific data mainly concerns the “experiment area” and areas of the scientific production (level 3) of the evaluation infrastructure. Regarding the experimental evaluation and the scientific production area, the conceptual model of DIRECT has been mapped into an RDF model and adopted for enriching and sharing the data produced by the evaluation activities.

In Fig. 3 we can see the classes and properties of the experiment area as reported and described in Silvello et al. (2017). Please note that the IMS namespace in this case indicates that all the class and property names are defined within the DIRECT workspace; this enables the distinction with other classes and properties in the LOD cloud which may have the same denomination, but of course different namespace. The area shown in the figure is central to the DIRECT infrastructure and it is connected to the most important resources for the evaluation activities.

Hence, we focus on this to present the semantic model we designed, even though it encompasses almost all the areas described above.

The experiment area can be divided into two main parts: one comprising the `Run`, `Track` and `Evaluation Activity` classes modeling the experiments and the other one comprising the `Quality Parameter`, `Measurement`, `Measure`, `Descriptive Statistic` and `Statistic` classes modeling the evaluation of the experiments.

The first part allows us to model an evaluation campaign composed of several runs submitted to a track which is part of an evaluation activity. The second part allows us to model the measurements and the descriptive statistics calculated from the runs and it is built following the model of quality for *Digital Library (DL)* defined by the DELOS Reference Model (Candela et al. 2007) which is a high-level conceptual framework that aims at capturing significant entities and their relationships within the digital library universe with the goal of developing more robust models of it; we extended the DELOS quality model and we mapped it into an RDF model. A `Quality Parameter` is a `Resource` that indicates, or is linked to, performance or fulfilment of requirements by another `Resource`. A `Quality Parameter` is evaluated by a `Measurement`, is measured by a `Measure` assigned according to the `Measurement`, and expresses the assessment of a `User`. With respect to the definition provided by the *International Organization for Standardization (ISO)*, we can note that: the “set of inherent characteristics” corresponds to the pair (`Resource`, `Quality Parameter`); the “degree of fulfillment” fits in with the pair (`Measurement`, `Measure`); finally, the “requirements” are taken into consideration by the assessment expressed by a `User`.

`Quality Parameters` allow us to express the different facets of evaluation. In this model, each `Quality Parameter` is itself a `Resource` and inherits all its characteristics, such as, for example, the property of having a unique identifier. `Quality Parameters` provide information about how, and how well, a resource performs with respect to some viewpoint. They express the assessment of a `User` about the `Resource` under examination. They can be evaluated according to different `Measurements`, which provide alternative procedures for assessing different aspects of a `Quality Parameter` and assigning it a value, i.e. a `Measure`. Finally, a `Quality Parameter` can be enriched with metadata and annotations. In particular, the former can provide useful information about the provenance of a `Quality Parameter`, while the latter can offer the possibility to add comments about a `Quality Parameter`, interpreting the obtained values, and proposing actions to improve it.

One of the main `Quality Parameters` in relation to an information retrieval system is its effectiveness, meant as its capability to answer user information needs with relevant items. This `Quality Parameter` can be evaluated according to many different `Measurements`, such as precision and recall (Salton and McGill 1983). The actual values for precision and recall are `Measures` and are usually

computed using standard tools, such as `trec_eval`,<sup>8</sup> which are Users, but in this case not human ones.

The `DescriptiveStatistic` class models the possibility of associate statistical analyses to the measurements; for instance, a classical descriptive statistic in IR is *Mean Average Precision (MAP)* which is the mean over all the topics of a run of the *Average Precision (AP)* measurement which is calculated topic by topic.

The described RDF model has been realized and implemented in the `DIRECT` system. This allows for accessing the experimental evaluation data enriched by the expert profiles that are created by means of the techniques that will be described in the next sections. This system is called `LOD-DIRECT` and it is accessible at the URL: <http://lod-direct.dei.unipd.it/>.

The data currently available include the contributions produced by the CLEF evaluation activities, the authors of the contributions, information about CLEF tracks and tasks, provenance events and the above described measures. Furthermore, this data has been enriched with expert profiles and topics which are available as linked data as well.

`LOD-DIRECT` serializes and allows access to the defined resources in several different formats such as XML, JSON, RDF+XML, Turtle<sup>9</sup> and Notation3 (n3).<sup>10</sup>

`LOD-DIRECT` comes with a fine-grained access control infrastructure which monitors the access to the various resources and functionalities offered by the system. Depending on the operation requested, it performs authentication and authorization.

The access control policies can be dynamically configured and changed over time by defining roles, i.e., groups of users, entitled to perform given operations. This allows institutions to define and put in place their own rules in a flexible way according to their internal organization and working practices. The access control infrastructure allows us to manage the experimental data which cannot be publicly shared such as log files coming from search engine companies.

## 4.1 Use Case

In Fig. 4 we can see an example of an RDF graph showing how `LOD-DIRECT` models topics, author profiles, measures and papers. This use case is taken from Silvello et al. (2017).

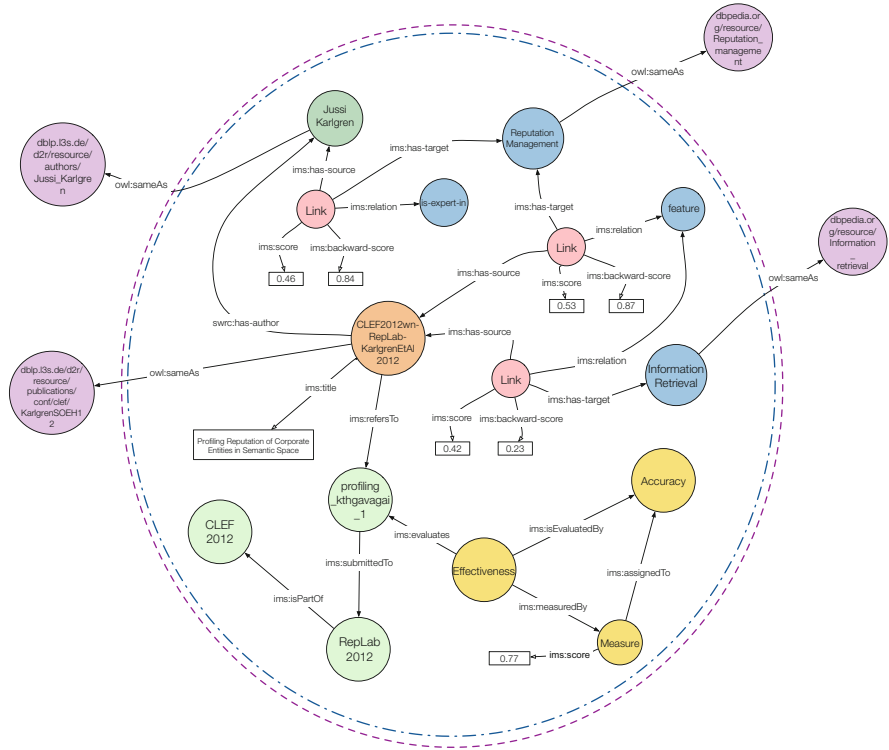
We can see the relationship between a contribution and an author enriched by expertise topics, expert profiles and connections to the LOD cloud. In this figure, we focus on the author (*Jussi Karlgren*) and the contribution (*KarlgrenEtAl-CLEF2012*). Here, there are two main topics, “reputation management” and “infor-

---

<sup>8</sup>[http://trec.nist.gov/trec\\_eval/](http://trec.nist.gov/trec_eval/).

<sup>9</sup><http://www.w3.org/TR/turtle/>.

<sup>10</sup><http://www.w3.org/TeamSubmission/n3/>.



**Fig. 4** An example of an RDF graph showing how expertise topics and expert profiles are used for enriching IR experimental data (Silvello et al. 2017)

mation retrieval”, which are related to the *KarlgrenEtAl-CLEF2012* contribution. We can see that *KarlgrenEtAl-CLEF2012* is featured by “reputation management” with a score of 0.53 and by “information retrieval” with 0.42, meaning that both these topics are subjects of the contribution; the scores give a measure of how much this contribution is about a specific topic. We can also see that the paper at hand presents the results for the RepLab 2012 track at CLEF 2012 where Gavagai obtained an accuracy of 0.77.

From this use case we see how LOD-DIRECT models the relationships between papers, authors, topics, measures and evaluation campaigns.

## 5 Related Work

A crucial question in IR is how to ensure the best exploitation and interpretation of the valuable scientific data employed and produced by the experimental evaluation. To the best of our knowledge, DIRECT is the most comprehensive tool for managing

all the aspects of the IR evaluation methodology, the experimental data produced and the connected scientific contributions.

There are other projects with similar goals but with a narrower scope. One is the *Open Relevance Project (ORP)*<sup>11</sup> which is a “small Apache Lucene sub-project aimed at making materials for doing relevance testing for Information Retrieval, Machine Learning and Natural Language Processing into open source”; the goal of this project is to connect specific elements of the evaluation methodology—e.g. experimental collections, relevance judgments and queries—with the Apache Lucene environment in order to ease the work of developers and users. Unfortunately, the project was discontinued in 2014. Moreover, ORP neither considers all the aspects of the evaluation process such as the organization of an evaluation campaign in tracks and tasks or the management of the experiments submitted by the participants to a campaign, nor takes into account the scientific production connected to the experimental data which is vital for the enrichment of the data themselves as well as for the definition of expert profiles.

Another relevant project is [EvaluatIR.org](http://EvaluatIR.org)<sup>12</sup> (Armstrong et al. 2009) which is focused on the management and comparison of IR experiments. It does not model the whole evaluation workflow and it acts more as a repository of experimental data rather than as an information management system for curating and enriching them.

There are other efforts carried out by the IR community which are connected to DIRECT, even though they have different purposes. One relevant example is the TIRA (TIRA Integrated Research Architecture) Web service (Gollub et al. 2012), which aims at publishing IR experiments as a service; this framework does not take into account the whole evaluation process as DIRECT does and it is more focused on modeling and making available “executable experiments”, which is out of the scope of DIRECT. Another relevant system is RETRIEVAL (Ioannakis et al. 2018); this is a web-based performance evaluation platform providing information visualization and integrated information retrieval for the evaluation of IR system. This system has some overlapping features with DIRECT, but it mainly focuses on the evaluation of IR systems rather than on the management of the data produced by evaluation campaigns and the management of the IR evaluation workflow.

## 6 Discussion

The DIRECT infrastructure effectively supports the management and curation of the data produced during an evaluation campaign. DIRECT has been used since 2005 for managing and providing access to CLEF experimental evaluation data. Over these years, the system has been extended and revised according to the needs and requirements of the community. Currently, DIRECT handles about 35 million

---

<sup>11</sup><https://lucene.apache.org/openrelevance/>.

<sup>12</sup><http://wice.csse.unimelb.edu.au:15000/evalweb/ireval/>.

documents, more than 13,000 topics, around four million relevance judgments, about 5000 experiments and 20 million measures. This data has been used by more than 1600 researchers from more than 75 countries world-wide.

Thanks to the expertise we have acquired in designing and developing it, we can now say that it would be preferable to have two distinct infrastructures rather than a single one:

- one to manage all those activities which are needed to run a cycle of an evaluation campaign;
- one for the long term preservation and curation of the information produced by the various evaluation campaign cycles over time.

In fact, DIRECT solves two different problems at the same time: those related to the management of the evaluation campaign cycles and those related to the archiving, preservation and curation of the experimental data produced by evaluation campaigns. However, these two kinds of activities are very different and managing them with a single infrastructure adds sizeable complexity to its design and implementation. On the other hand, if two distinct infrastructures were to be designed and implemented, each of them would be focused on a set of more homogeneous activities, resulting in simpler and more effective infrastructures for each specific objective. The results that are collected over time for each individual instance of an evaluation campaign could be used, for example, for activities of data analysis transversal to various periodic evaluation initiatives.

We have considered the possibility of developing two different infrastructures, because we believe that this effort would be extremely useful for the long-term development of the IR area. But developing two distinct infrastructures of this type would involve a significant investment of human and financial resources. Unfortunately, even if there is widespread agreement on the importance of experimental data, this kind of activity is not yet considered mainstream by the IR community. Therefore, to really value the effort and resources needed to implement such infrastructures, the IR community should better acknowledge the scientific value of such endeavours and should conduct them in a coordinated way so as to distribute the effort over different research groups and to produce a coordinated collection of scientific data that is at the same time curated, citable and freely available over the years for future scientific research.

**Acknowledgements** The results we have presented have mostly originated in the context of the research activities of the *Information Management System (IMS)* research group of the Department of Information Engineering of the University of Padua, Italy, but they have benefitted from the collaboration and the support of many experts, in particular of Carol Peters of ISTI, CNR, Pisa, Italy, and of Donna Harman of NIST, USA, to whom our sincere thanks are given. The research activities have been supported by the financial support of different European projects, namely DELOS (FP6 NoE, 2004–2007, Contract n. G038-507618), TrebleCLEF (FP7 CA, 2008–2009, Contract n. 215231), and PROMISE (FP7 NoE, 2010–2013, Contract n. 258191).

We are most grateful to our referees for their very helpful comments.

## References

- Agosti M, Di Nunzio GM, Ferro N (2007a) A proposal to extend and enrich the scientific data curation of evaluation campaigns. In: Sakay T, Sanderson M, Evans DK (eds) Proceedings of the 1st international workshop on evaluating information access (EVIA 2007). National Institute of Informatics, Tokyo, pp 62–73
- Agosti M, Di Nunzio GM, Ferro N (2007b) Scientific data of an evaluation campaign: do we properly deal with them? In: Peters C, Clough P, Gey FC, Karlgren J, Magnini B, Oard DW, de Rijke M, Stempfhuber M (eds) Evaluation of multilingual and multi-modal information retrieval: seventh workshop of the cross-language evaluation forum (CLEF 2006). Revised selected papers. Lecture notes in computer science (LNCS), vol 4730. Springer, Heidelberg, pp 11–20
- Agosti M, Di Nunzio GM, Ferro N (2007c) The importance of scientific data curation for evaluation campaigns. In: Thanos C, Borri F, Candela L (eds) Digital libraries: research and development. First international DELOS conference. Revised selected papers. Lecture notes in computer science (LNCS), vol 4877. Springer, Heidelberg, pp 157–166
- Agosti M, Ferro N, Thanos C (2009) DESIRE 2011: first international workshop on data infrastructures for supporting information retrieval evaluation. In: Ounis I, Ruthven I, Berendt B, de Vries AP, Wenfei F (eds) Proceedings of the 20th international conference on information and knowledge management (CIKM 2011). ACM Press, New York, pp 2631–2632
- Agosti M, Di Buccio E, Ferro N, Masiero I, Peruzzo S, Silvello G (2012) DIRECTIONS: design and specification of an IR evaluation infrastructure. In: Catarci T, Forner P, Hiemstra D, Peñas A, Santucci G (eds) Information access evaluation. Multilinguality, multimodality, and visual analytics. Proceedings of the third international conference of the CLEF initiative (CLEF 2012). Lecture notes in computer science (LNCS), vol 7488. Springer, Heidelberg, pp 88–99
- Agosti M, Fuhr N, Toms E, Vakkari P (2013) Evaluation methodologies in information retrieval (dagstuhl seminar 13441). *Dagstuhl Rep* 3(10):92–126
- Agosti M, Fuhr N, Toms EG, Vakkari P (2014) Evaluation methodologies in information retrieval Dagstuhl seminar 13441. *SIGIR Forum* 48(1):36–41. <https://doi.org/10.1145/2641383.2641390>
- Allan J, Aslam J, Azzopardi L, Belkin N, Borlund P, Bruza P, Callan J, Carman C, Clarke M, Craswell N, Croft WB, Culpepper JS, Diaz F, Dumais S, Ferro N, Geva S, Gonzalo J, Hawking D, Järvelin K, Jones G, Jones R, Kamps J, Kando N, Kanoulous E, Karlgren J, Kelly D, Lease M, Lin J, Mizzaro S, Moffat A, Murdock V, Oard DW, de Rijke M, Sakai T, Sanderson M, Scholer F, Si L, Thom J, Thomas P, Trotman A, Turpin A, de Vries AP, Webber W, Zhang X, Zhang Y (2012) Frontiers, challenges, and opportunities for information retrieval – report from SWIRL 2012, the second strategic workshop on information retrieval in Lorne, February 2012. *SIGIR Forum* 46(1):2–32
- Armstrong TG, Moffat A, Webber W, Zobel J (2009) EvaluatIR: an online tool for evaluating and comparing IR systems. In: Allan J, Aslam JA, Sanderson M, Zhai C, Zobel J (eds) Proceedings of the 32nd annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2009). ACM Press, New York, p 833
- Borgman CL (2015) Big data, little data, no data. MIT Press, Cambridge
- Bowers S (2012) Scientific workflow, provenance, and data modeling challenges and approaches. *J Data Semant* 1(1):19–30. <https://doi.org/10.1007/s13740-012-0004-y>
- Buneman P, Khanna S, Tan WC (2000) Data provenance: some basic issues. In: Kapoor S, Prasad S (eds) Foundations of software technology and theoretical computer science, 20th conference, FST TCS 2000 New Delhi, India, December 13–15, 2000, Proceedings. Lecture notes in computer science, vol 1974. Springer, Berlin, pp 87–93. [https://doi.org/10.1007/3-540-44450-5\\_6](https://doi.org/10.1007/3-540-44450-5_6)
- Candela L, Castelli D, Ferro N, Ioannidis Y, Koutrika G, Meghini C, Pagano P, Ross S, Soergel D, Agosti M, Dobрева M, Katifori V, Schuldt H (2007) The DELOS digital library reference model. Foundations for digital libraries. ISTI-CNR at Gruppo ALI, Pisa, Italy. <https://tinyurl.com/y7fxsz2d>



- Cleverdon CW (1997) The cranfield tests on index languages devices. In: Spärck Jones K, Willett P (eds) Readings in information retrieval. Morgan Kaufmann Publisher, San Francisco, pp 47–60
- Davidson SB, Buneman P, Deutch D, Milo T, Silvello G (2017) Data citation: a computational challenge. In: Sallinger E, den Bussche JV, Geerts F (eds) Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI symposium on principles of database systems, PODS 2017, Chicago, IL, USA, May 14–19, 2017. ACM, New York, pp 1–4. <https://doi.org/10.1145/3034786.3056123>
- Di Nunzio GM, Ferro N (2005) DIRECT: a distributed tool for information retrieval evaluation campaigns. In: Ioannidis Y, Schek HJ, Weikum G (eds) Proceedings of the 8th DELOS thematic workshop on future digital library management systems: system architecture and information access, pp 58–63
- Ferro N (2017) Reproducibility challenges in information retrieval evaluation. *ACM J Data Inf Qual* 8(2):8:1–8:4. <https://doi.org/10.1145/3020206>
- Ferro N, Hanbury A, Müller H, Santucci G (2011) Harnessing the scientific data produced by the experimental evaluation of search engines and information access systems. *Proc Comput Sci* 4:740–749
- Forner P, Bentivogli L, Braschler M, Choukri K, Ferro N, Hanbury A, Karlgren J, Müller H (2013) PROMISE technology transfer day: spreading the word on information access evaluation at an industrial event. *SIGIR Forum* 47(1):53–58
- Gollub T, Stein B, Burrows S, Hoppe D (2012) TIRA: configuring, executing, and disseminating information retrieval experiments. In: Hameurlain A, Tjoa AM, Wagner RR (eds) 23rd international workshop on database and expert systems applications, DEXA 2012, Vienna, Austria, September 3–7, 2012. IEEE Computer Society, Washington, pp 151–155
- Gray AJG, Groth P, Loizou A, Askjaer S, Brenninkmeijer CYA, Burger K, Chichester C, Evelo CTA, Goble CA, Harland L, Pettifer S, Thompson M, Waagmeester A, Williams AJ (2014) Applying linked data approaches to pharmacology. Architectural decisions and implementation. *Seman Web* 5(2):101–113
- Harman DK (ed) (1995) The fourth Text REtrieval Conference (TREC-4), National Institute of Standards and Technology (NIST), Special Publication 500–236, Washington, USA. [http://trec.nist.gov/pubs/trec4/t4\\_proceedings.html](http://trec.nist.gov/pubs/trec4/t4_proceedings.html)
- Harman DK (2011) Information retrieval evaluation. Morgan & Claypool Publishers, San Rafael
- Heath T, Bizer C (2011) Linked data: evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology. Morgan & Claypool Publishers, San Rafael
- Ioannakis G, Koutsoudis A, Pratikakis I, Chamzas C (2018) RETRIEVAL – an online performance evaluation tool for information retrieval methods. *IEEE Trans Multimedia* 20(1):119–127. <https://doi.org/10.1109/TMM.2017.2716193>
- Robertson SE (2008) On the history of evaluation in IR. *J Inf Sci* 34(4):439–456. <https://doi.org/10.1177/0165551507086989>
- Salton G, McGill MJ (1983) Introduction to modern information retrieval. McGraw-Hill, New York
- Sanderson M (2010) Test collection based evaluation of information retrieval systems. *Found Trends Inf Retr* 4(4):247–375
- Silvello G (2017) Theory and practice of data citation. *J Assoc Inf Sci Technol* 69:6
- Silvello G, Bordea G, Ferro N, Buitelaar P, Bogers T (2017) Semantic representation and enrichment of information retrieval experimental data. *Int J Digit Libr* 18(2):145–172
- Spärck Jones K, Bates RG (1977) Report on a design study for the ‘ideal’ information retrieval test collection. British Library Research and Development Report 5428, University Computer Laboratory, Cambridge
- Spärck Jones K, van Rijsbergen CJ (1975) Report on the need for and provision of an ‘ideal’ information retrieval test collection. British Library Research and Development Report 5266, University Computer Laboratory, Cambridge
- Voorhees EM (2002) The philosophy of information retrieval evaluation. In: Peters C, Braschler M, Gonzalo J, Kluck M (eds) Evaluation of cross-language information retrieval systems: second workshop of the cross-language evaluation forum (CLEF 2001) revised papers. Lecture notes in computer science (LNCS), vol 2406. Springer, Heidelberg, pp 355–370

- Voorhees EM (2007) TREC: continuing information retrieval's tradition of experimentation. *Commun ACM* 50(11):51–54
- Voorhees EM, Harman DK (2005) TREC: experiment and evaluation in information retrieval. The MIT Press, Cambridge
- W3C (2004) Resource description framework (RDF): concepts and abstract syntax – W3C recommendation 10 February 2004. <https://www.w3.org/TR/rdf-concepts/>
- Zapilko B, Schaible J, Mayr P, Mathiak B (2013) TheSoz: a SKOS representation of the thesaurus for the social sciences. *Seman Web* 4(3):257–263. <https://doi.org/10.3233/SW-2012-0081>
- Zobel J, Webber W, Sanderson M, Moffat A (2011) Principles for robust evaluation infrastructure. In: Proceedings of the workshop on data infrastructures for supporting information retrieval evaluation (DESIRE 2011), pp 3–6

# TIRA Integrated Research Architecture



Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein

**Abstract** Data and software are immaterial. Scientists in computer science hence have the unique chance to let other scientists easily reproduce their findings. Similarly, and with the same ease, the organization of shared tasks, i.e., the collaborative search for new algorithms given a predefined problem, is possible. Experience shows that the potential of reproducibility is hardly tapped in either case. Based on this observation, and driven by the ambitious goal to find the best solutions for certain problems in our research field, we have been developing the TIRA Integrated Research Architecture. Within TIRA, the reproducibility requirement got top priority right from the start. This chapter introduces the platform, its design requirements, its workflows from both the participants' and the organizers' perspectives, alongside a report on user experience and usage scenarios.

## 1 Introduction

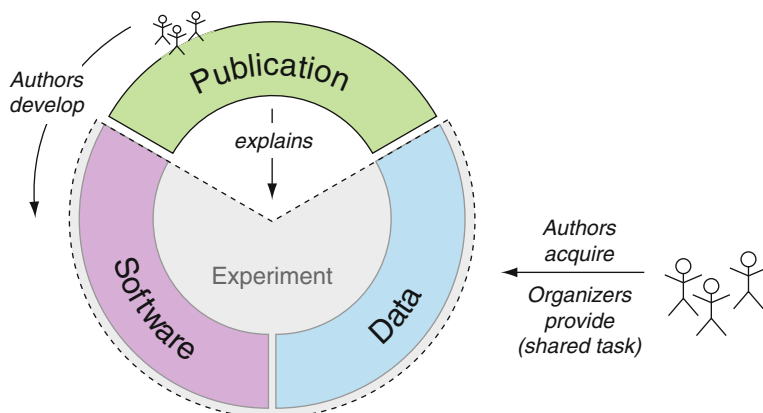
Computer science, when focusing on its data- and software-driven branches, is probably the only scientific discipline where the subject of research (the data) and its result (the software) can be copied, bundled, and shipped at virtually no extra cost—beyond what it took to acquire and create them respectively. The reproducibility of a computer science paper is greatly improved if the data and software underlying

---

M. Potthast (✉)  
Leipzig University, Leipzig, Germany  
e-mail: [martin.potthast@uni-leipzig.de](mailto:martin.potthast@uni-leipzig.de)

T. Gollub · B. Stein  
Bauhaus-Universität Weimar, Weimar, Germany  
e-mail: [tim.gollub@uni-weimar.de](mailto:tim.gollub@uni-weimar.de); [benno.stein@uni-weimar.de](mailto:benno.stein@uni-weimar.de)

M. Wiegmann  
Bauhaus-Universität Weimar, Weimar, Germany  
German Aerospace Center (DLR), Jena, Germany  
e-mail: [matti.wiegmann@uni-weimar.de](mailto:matti.wiegmann@uni-weimar.de); [matti.wiegmann@dlr.de](mailto:matti.wiegmann@dlr.de)



**Fig. 1** The three elements of reproducibility in computer science: publication, data, software. The two latter form the experiment. The software is developed by the authors of the publication, while the data may also be provided by a third party such as the organizers of a shared task

its experiments are available for the community. The publication (the paper) closes the circle by supplying motivation for the tackled problem, high-level descriptions of the courses of action taken, interpretation of the results obtained, and theories derived from observations made; see Fig. 1 for an illustration. However, the current practice in computer science differs: data and software of an experimental setting are typically not published, although they are the only tangible evidence of the claims made in a paper. From the outside it may look odd that scientists do not expose this evidence to third-party verification, thereby effectively reducing their papers to the level of anecdotes, until someone else comes along and double-checks.

It should be noted that the current economies of science do not enforce the reproducibility ideal (Stodden 2010); the commonly accepted measures for scientific excellence do not count movable assets other than papers. Any extra effort spent on data and software, despite increasing a publication’s impact and furthering one’s reputation, does not yield sufficient returns compared to moving on to the next topic or task. Not sharing data and software, however, poses an impediment to scientific progress: when someone decides to work on a task which has been tackled at least once before—effectively rendering it a shared task—they must take into account the workload required to reproduce missing assets to compare their own approach with those proposed earlier. The effort spent here is testimony to a scientist’s diligence, but many use only a small number of approaches for comparison, and sometimes none at all. Moreover, it can be difficult to avoid biasing experimental results while juggling the conflict of interest between optimizing one’s own approach versus those of third parties. Yet, even organized shared task events currently provide only part of the solution: while benchmark datasets are shared, software is typically not.

The chapter in hand introduces the TIRA Integrated Research Architecture, TIRA, a modularized platform for shared tasks. Section 2 provides background on computer science reproducibility and outlines our understanding of how shared

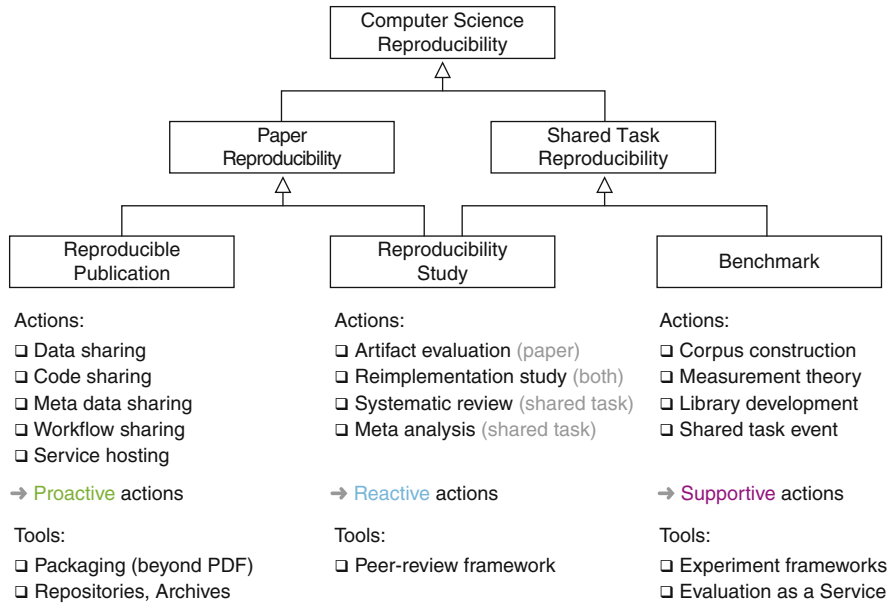


Fig. 2 Taxonomy of actions towards improving reproducibility in computer science

tasks emerge in computer science from which we derive requirements of shared task platforms. Section 3 introduces TIRA from the perspectives of participants and organizers, Sect. 4 generalizes the shared task paradigm from supporting software submission to data submission, and Sect. 5 reports how users have experienced the prototype.

## 2 Untangling the Reproducibility Thicket

Many initiatives across computer science seek to improve the situation with both organizational and technological means, aiming at easier sharing, citability of data and software, and better reproducibility. So far we have collected about 90 initiatives most of which are still active. Since a review of all of these initiatives and the related tools by far exceeds the scope and space limitations of this chapter, we have organized them into a taxonomy; see Fig. 2. Our high-level overview is oriented at the actions they have taken—which, by extension, every scientist can take for himself, too—to improve computer science reproducibility. TIRA is but one of these initiatives, providing a platform for shared task reproducibility. Otherwise, comprehensive overviews of relevant subsets of the existing initiatives have been recently compiled by Hanbury et al. (2015) and Freire et al. (2016).

## 2.1 Computer Science Reproducibility

Although reproducibility is one of the cornerstones of empirical science, its history in the empirical branches of computer science is comparably short. Claerbout and Karrenbach (1992), upon witnessing reproducibility problems within computational parts of their geophysics project, were among the first to force the issue by implementing rigorous guidelines for the project members, and publishing about them. Some followed their example and developed their own strategies (e.g., Donoho et al. 2009). However, the majority of computer scientists continued with business as usual, occasionally interrupted by some who denounced the lack of reproducibility when failing to verify published results by reimplementing (e.g., Pedersen 2008; Fokkens et al. 2013). Pioneered by Stodden et al., systematic research into computer science reproducibility started only recently,<sup>1</sup> in the wake of what has become known as science's reproducibility crisis,<sup>2</sup> which struck home particularly hard in the life sciences, and which brought the issue into focus for computer scientists as well.

Interestingly, the terminology used to describe efforts related to demonstrating, checking, or otherwise concerning the reproducibility of a piece of research is rather ill-defined to this day (Plesser 2018), riddled with misunderstandings and contradictory definitions. One of the many attempts to define reproducibility has been provided recently by the Association for Computing Machinery (ACM) as part of a new peer-review initiative called "artifact evaluation," which is poised to supplement traditional peer-review at conferences throughout computer science. The initiative distinguishes three levels of reproducibility<sup>3</sup>:

- Repeatability (Same team, Same experimental setup)

The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation.

- Replicability (Different team, Same experimental setup)

The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.

- Reproducibility (Different team, Different experimental setup)

---

<sup>1</sup><http://web.stanford.edu/~vcs/Papers.html>.

<sup>2</sup>[https://en.wikipedia.org/wiki/replication\\_crisis](https://en.wikipedia.org/wiki/replication_crisis).

<sup>3</sup><https://www.acm.org/publications/policies/artifact-review-badging>.

The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.

Although these definitions are not necessarily the best ones, in combination with Fig. 2, they are well-suited to delineate what can and what cannot be accomplished with certain reproducibility tools in general (and TIRA in particular). Used in isolation, the term “reproducibility” encompasses all of these aspects.

In general, we distinguish efforts that target the reproducibility of an individual paper from efforts that target the reproduction of *a set of papers* all of which tackling the same (shared) task. Of course, also papers of the latter type are unique in the sense that they employ tailored methods or pursue customized solution approaches. The distinction (individual paper versus shared task papers) emphasizes that the latter address well-defined tasks which have been tackled before and for which the authors need to compare their approach to those of their predecessors. While ensuring the reproducibility of individual papers involves striving for completeness in terms of the specific experimental setups, ensuring the reproducibility of shared tasks requires some form of abstraction and unification, disregarding approach-specific details.

The efforts related to individual paper reproducibility result from the following goal: Build a kind of self-contained, reproducible publication that, even in the absence of its authors, can be used to replicate its results from scratch (Fig. 2, bottom left box). Actions that authors can *proactively* take to render their publications reproducible may be subsumed under the motto “Share all research-relevant assets.” Tools that have been proposed for this purpose include packaging software, which compiles all assets into one publication package (a single file) which in turn can be shipped and published alongside the traditional PDF (e.g., ReProZip). In this regard, asset-specific repositories and archives have been established, providing long-term preservation for certain assets (e.g., Linguistic Data Consortium, European Language Resources Association, Zenodo) as well as citability to serve as further incentive for authors to make their publications reproducible.

The efforts that can be taken by publishers, conference organizers, and third party stakeholders to further incite authors to ensure the reproducibility of their papers are reproducibility studies (Fig. 2, bottom middle box). Subject to such studies can be individual papers or sets of papers about a shared task. Note that, by nature, such studies are *reactive*; i.e., they can be done only when authors have finished their papers and are about to publish them. A fairly recent approach is to consider reproducibility within peer-reviews, either by asking reviewers of the PDF version of a paper to judge the reproducibility of its underlying assets (usually without accessing them), or by introducing an additional, dedicated review cycle called artifact evaluation: the participating authors share their assets for review and are awarded badges of honor when the reviewers find them sufficient.

Apart from these gatekeeping actions, which cannot be arbitrarily scaled and hence will affect only a small portion of papers published, third parties can

undertake a reimplementing study of either one or a set of papers. In this vein, some conferences have introduced special tracks on reproducibility to render publishing such studies easier. Another approach to analyze the reproducibility of a set of papers studying the same task are systematic reviews and meta studies, which, even without reimplementing, can reveal systematic biases in experimental setups.

Last but not least, *supportive* efforts taken by leading scientists on an established shared task include the development of effective benchmarks and gold standards (Fig. 2, bottom right box). If a benchmark is accepted and adopted by a larger portion of the community around a shared task, it ensures comparability of all papers using it, and its adoption by newcomers is often enforced via peer-reviewing, eventually rendering the benchmark “self-propagating” as the number of papers using it increases. Specific actions that support the development of benchmarks are the creation of task-specific resources, such as corpora and evaluation datasets, inquiries into measurement theory with respect to a task, developing software libraries, and not least, the organization of shared task events where participants are invited to work on a given task for which the necessary resources are provided. It can be observed that shared task events always have been instrumental in both the creation of benchmarks and the coordination of evaluation activities. For decades, especially in the human language technologies, entire conferences have been organized, some hosting dozens of shared task events at a time: TREC, CLEF, NTCIR, FIRE, SemEval, MediaEval, the KDD Cup, or CoNLL’s shared task track, to name only some of the best-known ones. TIRA has been developed within the PAN lab on digital text forensics, hosted at the CLEF conference since 2010.

Considering the above definitions of the three reproducibility levels, it becomes clear that a tool that supports proactive actions will also improve repeatability and replicability, but not reproducibility. In fact, improving the former may have adverse effects on the latter: the more assets are made available, the smaller is the need to reproduce them independently from scratch. Independent reproduction then becomes a deliberate decision instead of an everyday task for those who wish to compare their approaches to the state of the art. Again, computer science may occupy a unique position compared to other disciplines: because of the ease with which its assets can be shipped, once they are available, they may be reused without second thought, this way propagating potential biases and errors encoded in them. This characteristic is different from other empirical sciences, where sharing experimental assets is infeasible (e.g., human test subjects), so that new claims need to be independently reproduced sufficiently often before being included into the scientific canon.

TIRA fits into this picture as follows. It is a platform that has been devised as a powerful tool to support the organization of shared task events. As its most salient feature we consider the ease by which software can be submitted and, in particular, maintained for future re-execution. TIRA operationalizes blind evaluation, a paradigm that is rarely applied in empirical computer science. The term refers to an evaluation process where the authors of a to-be-evaluated piece of software cannot access the test data and hence cannot (unwittingly) optimize their algorithm against it. For this purpose TIRA implements a kind of airlock for data through which the to-be-evaluated software has to pass. The software itself is packaged within a virtual



machine and hence can be archived in working condition. Given these concepts, TIRA facilitates repeatability and replicability in first place. In addition, TIRA also supports an important reproducibility aspect, namely, the execution of a TIRA-hosted software on newly-created datasets, i.e., datasets which were not available at the time of software creation. This is ensured by providing a unified software execution interface along with harmonized dataset formats of a shared task, so that drop-in replacements become possible. At the time of writing, there are not many cloud-based evaluation systems comparable to TIRA. Collectively, these systems implement the so-called evaluation as a service paradigm. An overview of such systems has been recently compiled by Hanbury et al. (2015).

## 2.2 *Shared Tasks: From Run Submission to Software Submission*

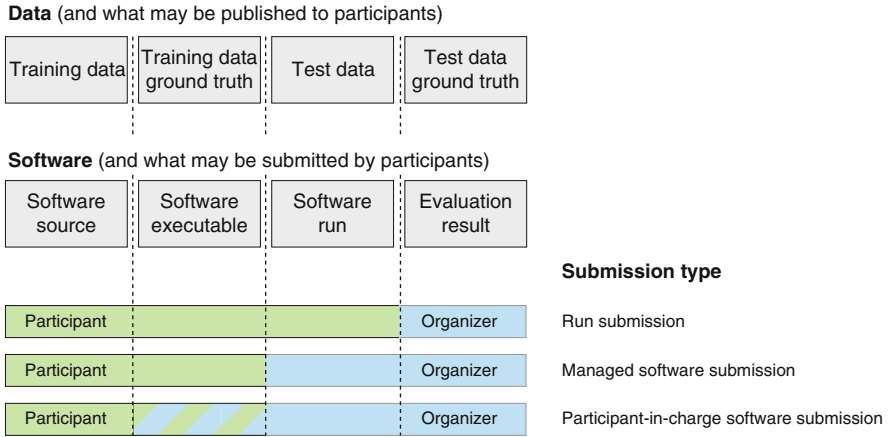
When two or more groups of scientists start working independently on the same task, it literally becomes a shared task, albeit an uncoordinated one.<sup>4</sup> The initiators or “inventors” of a shared task carve out the experimental setup, including evaluation resources and performance measures, and the followers may refine it or propose a new one. As a task’s followership grows, commonly accepted setups emerge and reviewers familiar with the task will hint at its proper evaluation. Eventually some stakeholder may organize an event around a shared task. The central goal of such an event is to compare latest algorithmic approaches to solving the task’s underlying problem in a controlled laboratory experiment.

A review of existing shared task events in the human language technologies revealed that such tasks have been almost unanimously organized in the same way; see Fig. 3 for an illustration. Task organizers prepare a dataset comprising problem instances, where parts of the dataset are published as training data (including the ground truth) and test data (without the ground truth) respectively. Task participants develop pieces of software that solve the task based on the training data and finally run their software on the test data. Within most shared tasks, the output of this final software run (called a *run* for short) is submitted to the organizers. The organizers, in turn, evaluate the submitted runs based on previously announced performance measures against the ground truth of the problem instances in the test data set.

To reach higher levels of automation and reproducibility, participants may submit their executable software, enabling the organizers to generate outputs by themselves, an approach we call “managed software submission.” A major obstacle to a widespread adoption of managed software submission in shared tasks is the shift of responsibility for a successful software execution. Submitted software is not necessarily free of errors—even more, experience shows that many participants submit their software prematurely, being convinced of its flawlessness. This fact

---

<sup>4</sup>The etymology of the term “shared task” is unclear; conceivably, it was coined to describe a special kind of conference track and was picked up into general use from there. We generalize the term, interpreting it literally as referring to any set of papers sharing a task.



**Fig. 3** From top to bottom: Task organizers develop a dataset from which certain parts are published to participants. The participants in turn develop software from which certain parts are submitted. The extent of what is published or submitted defines the submission type: run submission, managed software submission, or participant-in-charge software submission. The last submission type enables participants to submit, execute, and optimize their software, using an experiment platform (such as TIRA) provided at the organizer’s site

makes organizers unwillingly become part of the debugging process of each participant’s software, and the turnaround time to find and fix errors increases severely, especially when both parties are not working simultaneously (i.e., reside in different time zones). Failure on the part of organizers to run a submitted software, to check its output for errors of any kind (e.g., not every execution error results in a crash), and to give participants feedback in a timely manner may cause participants to miss submission deadlines. The risk of this happening is increased by the fact that many participants start working only just before a deadline, so that organizers have to handle all submissions at the same time. Besides, prolonged back-and-forth between participants and organizers caused by software errors bears a high potential for friction. As a result, organizers may come to the conclusion they have little to gain but trouble, whereas the benefits of software submissions, such as reproducibility, may be considered insufficient payback.

Our experience with managed software submissions at a shared task organized in 2013 is as follows (Gollub et al. 2013): to get the 58 pieces of software submitted running for evaluation, 1493 mails had to be exchanged in order to fix runtime errors. We postpone a more in-depth analysis of this mail exchange to Sect. 5. For now, suffice it to say that we were working hand-in-hand with participants, and that, surprisingly, participants affected by software bugs were not at all disgruntled about revisiting their software over and over again to fix them. From this experience, we derived a list of requirements (detailed in the next subsection) that a platform for shared tasks that implements software submission should fulfill. Most importantly, it must keep participants in charge of their software,

lifting responsibility from the organizers shoulders, and reducing the overhead to a bearable level for both sides.

We thus introduced a third kind of submission type, here called “participant-in-charge software submission,” providing a self-service evaluation to participants. Under this paradigm, the software is submitted to a cloud-based evaluation platform, which provides for a suitable runtime environment and manages the software. Crucially, it gives participants full control over their software and whether it works by providing runtime feedback, e.g., when run on the training data. At the request of participants, the software gets to access the test data, but that includes cutting off participant access to the software as well as the moderation of any runtime feedback by organizers. Though this approach is technically the most advanced, bringing about corresponding technical difficulties, it also comes along with appealing advantages: the software can be tested and optimized by the participants, as well as accessed, run, and archived for documentation and re-run purposes by the organizers.

### **2.3 Requirements Analysis**

Computer science is bustling with hundreds of uncoordinated shared tasks that offer potential for future growth into shared task events. The question remains if and how the process of nurturing a shared task from its uncoordinated beginnings into a reliable evaluation activity can be formalized (and hence simplified and accelerated), and what are the requirements for a platform to support this process. Based on our experience with organizing shared task events which invited run submissions and managed software submissions, we have derived the following list of requirements.

*Technological Compatibility* A key requirement of a platform is its compatibility with new technologies, which emerge at a rapid pace and which have been leading to a great diversity of technology stacks and software development environments. Every developer has their own technical preferences, and minimizing compatibility constraints will help to not alienate potential users.

*Setup Multiplicity* Since shared tasks emerge from day to day work in a lab, those who create the first experimental setup for a given task automatically take the role of an “organizer,” defining data formats and interfaces. While their influence on successive scientists and their experimental setups is strong, one cannot expect them to get their setup right the first time around. Growing understanding of how a task can be tackled influences how it should be evaluated and dedicated scientists will not adopt a setup that does not reflect the most recent understanding. Sticking to the first setup, or excessively amplifying its importance, will result in the rejection of a platform. To avoid such conflicts, a requirement is to allow for running multiple competing experimental setups simultaneously for the same task.

*Plugin Functionality* Once the experimental setup of a shared task has taken shape, i.e., when all interfaces are defined, various stakeholders will develop new resources to complement existing ones. A platform for shared task creation should support the submission of all kinds of resources and, besides using them for the development of new solutions to a task, also allow third parties to analyze them. In a nutshell, the control over resources has to be relinquished to the community. Resources for a shared task include datasets, performance measures, and visualizations of datasets as well as of results. In addition, more complex tasks will require the integration of task-specific tools, e.g., in the form of external web services. A corresponding plugin architecture should hence be available.

*Software Execution Layer* The integration of new resources to a shared task immediately raises the question as to what performance previously evaluated approaches will achieve on them. Current shared task events typically do not collect the software of participants, so that future comparative evaluations are restricted to the same experimental setup that has been used for the original event. A platform hence should allow the reproducibility of shared task events by inviting participants to submit their software, which, of course, must be maintained and kept in working condition. Under ideal conditions, previously submitted software can be re-evaluated as soon as new resources become available for a given task.

*Export and Service Layer* In addition to low integration barriers both for data and software, there should also be the possibility (for all users) to export resources and to run corresponding evaluations locally instead of on the platform itself. At the same time, submitted software should be exposed via APIs if their developers wish to share it, e.g., to foster the transfer of well-performing solutions to practical use.

*Governance Functionality* Even if all mentioned requirements regarding technological flexibility are fulfilled, the success of a shared task platform will eventually depend on its adoption and use in practice. The people involved in a shared task must be connected with each other as well as to third parties interested in making use of the developed software. Providing a social layer can help to establish shared task governance structures, allowing for coordinated task maintenance and development. Some stakeholders may take a leading role within the community, becoming what could be called a “shared task editor.” Unlike the organizers of traditional shared task events, who typically create the complete experimental setup required, editors can ensure interface compliance of new resources, lead the community regarding certain aspects of a shared task, and organize workshops where the recent developments can be discussed. Since shared task editorship can switch between community members, the ideal shared task platform should provide basic governance functionality.

*Security* Executing the software of malicious participants exposes one’s infrastructure and data to exploitation. Simultaneously, submitting one’s software to a shared task platform renders it open to theft should the organizer’s systems be insecure. Moreover, hosting valuable data on such a platform invites attempts at stealing it by exploiting vulnerabilities of the platform. Therefore, a platform for shared tasks must be built with a keen eye on security from the bottom up.

### 3 TIRA: An Architecture for Shared Tasks

This section reports on our efforts to build a platform for shared tasks, the TIRA Integrated Research Architecture (Gollub et al. 2012a,b). TIRA has been used to organize shared task events right from the start in 2012. To date, it has handled more than a dozen shared task events and hundreds of participants and their software submissions. With the TIRA prototype, we propose solutions to most of the aforementioned requirements for shared task platforms. In particular, TIRA allows for development environment diversity, untrusted software execution, it prevents data leakage, and supports error handling by participants. Public access to TIRA's web front end is available,<sup>5</sup> and its code base is shared open source.<sup>6</sup> In what follows, we overview TIRA's architecture, its two most salient contributions (the datalock and blind evaluation), and give a detailed view of the user interfaces and the workflow of participants and organizers to complete a shared task.

#### 3.1 Architecture Overview

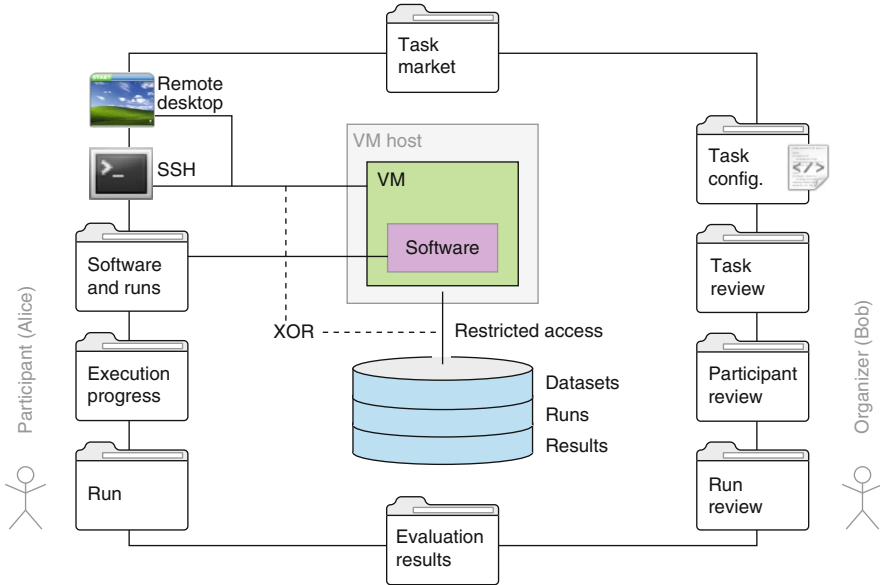
Figure 4 gives an overview of TIRA's basic architecture. TIRA's three main components include a web service hosting user interfaces for participants and organizers, host servers for virtual machines (VM), and a storage server. All three may be running on the same physical machine, but are typically distributed across a data center. The storage server serves as a central repository for evaluation datasets, runs of participant software, and evaluation results obtained from measuring performance by comparing runs with the ground truth of a task's datasets. Participant software is encapsulated by virtual machines, and typically each participant is assigned one virtual machine per task, whereas the resources allocated to each machine may vary.

For brevity, we omit a description of the technical details and the software stack making up TIRA right now. Because of the ongoing development on the prototype, the current software stack will likely be replaced by another one at some point in the future, while the main features and functions of TIRA remain unchanged. A better way to understand TIRA is to study its most salient features, and by tracing the workflow of a typical participant and that of a typical organizer. These workflows are implemented within TIRA's web interfaces, which allow participants to remotely control software execution and to collect runtime feedback, thus eliminating the need for organizers to intervene in fixing software execution errors. At the same time, the organizers are given a number of interfaces to watch over a shared task and its participants.

---

<sup>5</sup><http://www.tira.io>.

<sup>6</sup><http://www.github.io/tira-io>.



**Fig. 4** TIRA's interfaces for participants (left), organizers (right), and the public (top, bottom)

### 3.2 The Datalock: Preventing Data Leakage

Software submitted by participants of a shared task must be treated as untrusted software, i.e., as software that may include malicious code. Virtual machines have been used for years already in all kinds of cloud platforms that allow renting servers to host and execute software, secluding every customer's software from one another. We adopt the same principles for TIRA. However, unlike in typical cloud platforms, the software deployed to the virtual machines of participants is supposed to process datasets hosted on TIRA as well, in particular the test datasets of the shared tasks. It is important that the integrity of these datasets remains intact, regardless of which software processes it, and that the test datasets do not leak.

To ensure that test data are sufficiently protected, TIRA implements what can be described as an airlock for data, the datalock. Before a software is started, the virtual machine to which it has been deployed is passed into the datalock together with the test data, where it can process the data and have its output recorded. Otherwise, all communication channels to the outside are closed or tightly moderated. To pass into the datalock, the virtual machine is cloned or snapshotted, and the copy is disconnected from the internet so that no incoming or outgoing connections are possible. The test data are then mounted read-only to the virtual machine copy. Only if a machine has been successfully moved into the datalock, is the software executed. Disconnecting the copied virtual machine from the internet while a software is executed ensures that no data can be automatically sent to an unauthorized third

party and that participants have no access to the copy of their virtual machine during software execution. After the software terminates, the output is stored in TIRA's database as a run, and the virtual machine is automatically moved out of the datalock: this boils down to either deleting the clone or resetting the virtual machine to the time of the snapshot. Deleting the copy ensures that no information about the input data remains, be it in cache, in temporary files, or in purposefully hidden files. This way, the only communication channel left is the software's output, which is kept hidden from participants by default, unless a task organizer reviews the run and concludes that it does not reveal anything important about the test data.

With some reservations, the datalock allows for shared tasks to be organized on the basis of secret, sensitive, and proprietary data. It must be conceded, though, that the security of the datalock hinges on that of the operating systems, hypervisors, and other dependent libraries of TIRA. Especially regarding extremely sensitive data, such as medical data, it may be necessary to not give participants any feedback at all, since vital information may be encoded into even the narrowest communication channel. Conversely, the fewer feedback channels remain open during software execution, the more difficult and time-consuming it becomes for participants to become aware of any software errors, since a task organizer needs to intervene first. Nevertheless, for the vast majority of shared tasks that arise from private data, the concept of the datalock constitutes a first-time opportunity for their owners to give third parties access without sharing the data itself.

### ***3.3 Blind Evaluation for Shared Tasks***

In the vast majority of evaluations carried out in empirical computer science, the experimenters have direct access to the test data and its ground truth. The validity of computer science experiments builds on trusting the experimenter to ignore the test data and its ground truth until the to-be-evaluated software is ready (we may call this approach "pseudo-blind"). Similarly, the vast majority of shared task events are organized half-blind (the test data are shared with participants but the ground truth is withheld as depicted in Fig. 3). Ideally, however, an evaluation would be conducted blind, so that a scientist has access only to the training data but not to the test data, rendering it difficult to fine-tune a given approach against the latter.

Remaining ignorant of the test data is incredibly important for an evaluation, and not doing so may have a significant impact on the validity of evaluation results. In general, one can only trust that scientists do not spoil their experiments by implicitly or explicitly exploiting their knowledge of the test data. Within shared task events, another factor comes into play: shared task events are also competitions. Dependent on its prestige, winning a shared task event may come along with a lot of visibility, so that supplying participants with the test data up front bears risks of cheating, and mistakes that spoil the ground truth.

Together with participant-in-charge software submission, the datalock may give rise to a widespread adoption of blind evaluation in shared tasks. TIRA enforces

blind evaluation by default, as long as the allowance of the number of runs against the test data is kept small by organizers.

### 3.4 *Life of a Participant*

From the perspective of a participant (Alice, in the following), a software submission via TIRA happens within three steps: first, deployment of the software to a given virtual machine, second, configuration of the software for remote execution, and third, remote execution of the software on the available datasets. The interfaces on the left side of Fig. 4 are used for this purpose. They put Alice in charge of deploying her software, allowing her to evaluate her software and to obtain (moderated) runtime feedback. TIRA serves as a remote control for evaluation.

TIRA encapsulates Alice's software in a virtual machine that is set up once she registers for a shared task. As depicted in Fig. 4, Alice has two ways to access her virtual machine, namely a remote desktop connection and an SSH connection. Alice retains full administrative rights inside her virtual machine, so that she can set up her preferred development environment and deploy her software. To prevent misuse, virtual machines are not allowed to communicate with each other, and, their outgoing bandwidth is limited. By default, virtual machines have only restricted access to TIRA's database, so that only the training data of each task can be read. Once a software has been successfully deployed and tested manually, participants use TIRA's web interface to complete the second and third step outlined above.

For each participant of a shared task, TIRA serves a page for the respective virtual machine, the deployed software, and the software runs. After signing in with her account for the first time, Alice can configure the execution details of her software. Figure 5 shows Alice's software control page in a state after completed configuration and a few successfully executed runs:

*Virtual Machine* Overview of the virtual machine, including information about the operating system, RAM, CPUs, its running state, VM host, and connectivity. The virtual machine can be turned off at the click of a button either by sending a shutdown signal to the operating system, or by powering it off. Clicking "Add Software" creates a new software panel. Alice may deploy an arbitrary number of pieces of software for the shared task on her virtual machine, e.g., to compare different paradigms or variants of an approach at solving the task. Each software can be configured individually on the software control page.

*Software 1* Configuration of a software that has been deployed on the virtual machine. The software must be executable as a POSIX-compliant command. Mandatory parameters can be defined by organizers of the shared task. In this case, they include variables for input data and the output directory, and optionally for an input run (i.e., a previous run of one of Alice's pieces of software). Optionally, the working directory in which the program will be executed can be specified. Alice may adjust an existing software configuration and save its state, she may delete it, or



### Virtual Machine

**Operating System** Ubuntu (64 bit)  
**RAM** 4096MB  
**CPUs** 1  
**State** running (since 2014-06-22 09:00:00)  
**Sandbox state** publicly accessible  
**Host** example.com  
**SSH Port** 44401 [open](#)  
**RDP Port** 55501 [open](#)

[Add software](#) [Shutdown](#) [Power off](#)

### Software 1

**Command**   
The variables `$inputData` and `$inputRun` refer to the below parameters; the command must include the variable `$outputDir`. All of these variables will point to directories.

**Input data**

**Input run**   
Runs on test corpora are excluded from this list.

**Working directory**

[Save](#) [Delete](#) [Run](#)

### Evaluation

**Measures** precision, recall, accuracy

**Input run**   
Evaluator runs are excluded from this list.

[Run](#)

### Runs

| Software   | Run                 | Input data    | Input run           | Runtime  | Size | Actions   |
|------------|---------------------|---------------|---------------------|----------|------|---|
| evaluation | 2014-06-22-12-10-00 | test-data     | 2014-06-22-12-00-00 | 00:00:04 | 24K  | <a href="#">i</a> <a href="#">d</a> <a href="#">x</a> |
| software1  | 2014-06-22-12-00-00 | test-data     | none                | 00:01:54 | 2.2M | <a href="#">i</a> <a href="#">d</a> <a href="#">x</a> |
| software1  | 2014-06-22-11-00-00 | training-data | none                | 00:01:54 | 2.2M | <a href="#">i</a> <a href="#">d</a> <a href="#">x</a> |
| software1  | 2014-06-22-10-00-00 | training-data | none                | 00:00:30 | 1.1M | <a href="#">i</a> <a href="#">d</a> <a href="#">x</a> |

Fig. 5 TIRA’s web interface for participants to remotely control the execution of their software and to review their runs for a given shared task

she may proceed to execute the software. If Alice deletes a software it is not actually deleted on the server, but only hidden from view; rationale for this is to allow the reconstruction of Alice's actions for cheating prevention. The runs obtained from running a software are listed in the "Runs" panel.

*Evaluation* Runs an evaluation software on a given run. This is a special type of software provided by task organizers which processes an input run and outputs the results of the task's performance measures. Once Alice has finished her first successful run on a given input dataset, she uses this panel to evaluate it. The runs obtained from an evaluation software are also listed in the "Runs" panel.

*Runs* List of runs that have been obtained either from running a software or from running an evaluation. The table lists run details including software, timestamp, input data, input run, runtime, size on disk, and further actions that can be taken. The colorization indicates a run's status with respect to its success, where red indicates severe errors, yellow indicates warnings, green indicates complete success, and white indicates that the run has not yet been reviewed. Runs may be checked automatically for validity with the shared task's expected output format, and they may be reviewed manually by organizers. Actions that can be taken on each run include viewing more details (the blue *i*-icon), downloading it (the black arrow down), and deleting it (the red x). It is here where Alice first encounters the limitations that TIRA imposes for runs on test data: all test datasets are by default hidden from participants. TIRA prevents Alice from downloading runs on test datasets (the download action shown in gray is inactive) to prevent a malicious software from outputting the data itself instead of outputting the data that is valid for a given shared task. Even runtime and size information are initially hidden, unless a task organizer decides to unblind them.

The software control page does not display all of the aforementioned panels immediately, but only after Alice has completed the necessary steps. At first, it only shows the virtual machine panel; then, once Alice clicks on "Add Software", a software panel appears; and finally, once Alice runs her software for the first time, the evaluation panel and the runs panel are added after the run is completed. While a software is running, the software control page is replaced with the software progress monitoring page which is divided into two panels, as exemplified in Fig. 6:

*Virtual Machine* Just as on the software control page, the virtual machine panel shows the current state of Alice's virtual machine while the software is running. Before a software is started, the virtual machine is moved into the datalock as described above. This process may take some time, so that the intermediate states of the virtual machine are indicated in this view. The port flags indicate to Alice that access from the outside to her virtual machine has been disabled. Only if a machine has been successfully moved into the datalock, is the software executed. While a software is running, the buttons to add a software configuration panel as well as those to shutdown or power off the virtual machine are deactivated so that the running software is not interrupted accidentally. After the software terminates, the

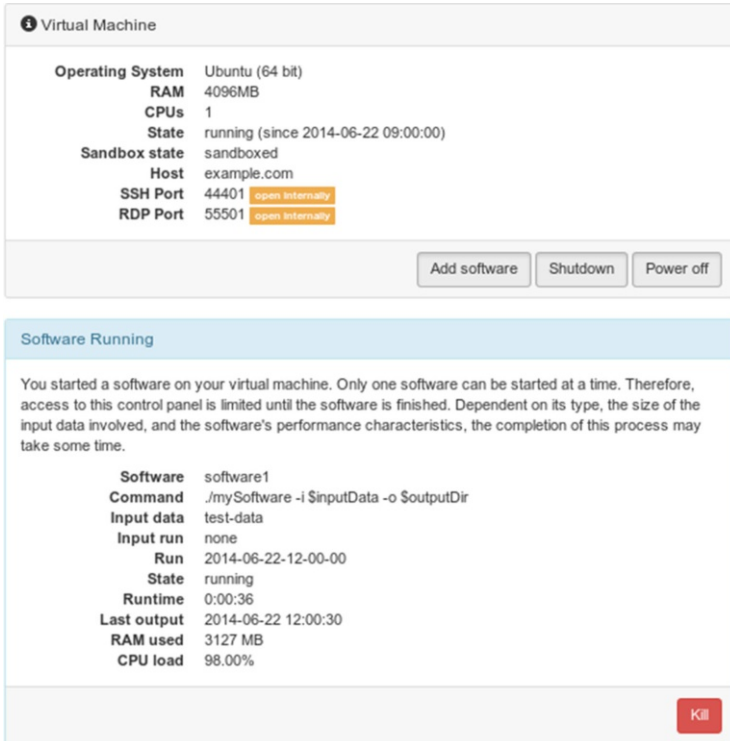


Fig. 6 TIRA’s web interface to monitor the progress of a running software

output is stored in TIRA’s database as a run, and the virtual machine is automatically moved out of the datalock.

*Software Running* Overview of a running software, including the software’s ID, the executed command, the parameters, the run ID, and the running state. Moreover, the current runtime, the time of the last write access to the output directory, the currently used RAM, and the CPU load are displayed and updated periodically. This way, Alice has a way of making sure her software is still working. If, for any reason, Alice wishes to kill her software before it terminates by itself, she may click on the “Kill” button. Before the software is killed, its output up to this point is stored in TIRA’s database as an incomplete run for later inspection.

After her run has completed and the virtual machine has been moved out of the datalock, Alice’s browser shows the software control page again as in Fig. 5. The new run appears in the runs table. To make sure the run was successful, Alice clicks on the *i*-icon which redirects her to a run details page for the run in question, as shown in Fig. 7a. The details shown about a run are as follows:

*Overview* Details about the run, including the software that was used, the run ID, parameters, whether the run can be downloaded, runtime details, its size, and the

(a)

←
Run Details

Overview

|                        |  |
|------------------------|--|
| <b>Software</b>        | software1  |
| <b>Run</b>             | 2014-06-22-12-00-00  |
| <b>Input data</b>      | test-data  |
| <b>Input run</b>       | none   |
| <b>Downloadable</b>    | false  |
| <b>Runtime</b>         | 00:01:54 (hh:mm:ss)  |
| <b>Runtime details</b> | 96.79user 8.79system 1:54.81elapsed 91%CPU (0avgtext+0avgdata<br>202016maxresident)k 224inputs+4160outputs (0major+14449minor)pagefaults<br>0swaps |
| <b>Size</b>            | 2.2M (154442 bytes)  |
| <b>Lines</b>           | 0  |
| <b>Files</b>           | 518  |
| <b>Directories</b>     | 1  |

Review

|                 |                                     |
|-----------------|-------------------------------------|
| <b>Reviewer</b> | Bob                                 |
| <b>Errors</b>   | None. This run seems to be alright. |

Stdout

```
[...]t516.xml
Processing input517.xml
Writing output517.xml
Processing input518.xml
Writing output518.xml

Note: The output of software that is run against test data is shortened to
its last 100 chars.
```

Stderr

File list

```
test-data/alice/2014-06-22-12-00-00/output
├── [ 90] output1.xml
├── [ 257] output2.xml
├── [ 90] output517.xml
├── [ 255] output518.xml
└── 0 directories, 518 files
```

(b)

Stdout

```
python shared-task-evaluation.py -i alice/2014-06-22-12-00-00/output -t
test-data -o /tmp/2014-06-22-12-10-00/output/evaluation.txt

"precision": "XXX"
"recall": "XXX"

Note: The output of evaluation runs on test corpora is blinded by default.
A task moderator will decide whether to make the results visible.
```

Stderr

**Fig. 7** TIRA's web interfaces for participants to review runs. (a) Details page of a software run. (b) Excerpt of the details page of an evaluation run

numbers of lines, files, and directories found. Whether the run can be downloaded depends on whether the input data was a test dataset or not. As outlined above, runs on test datasets, by default, cannot be downloaded to foreclose data leakage. Besides the runtime, more in-depth runtime details are given, so that Alice can judge whether her software made good use of the hardware resources available to the virtual machine. For example, if she finds there are many page faults or even swaps, this indicates the software uses too much memory. The size and numbers of lines, files, and directories provide quantitative feedback to quickly verify output sanity, whereas it depends on the task which of these values is most illuminating.

*Review* Review of this run provided by both automatic validation and organizers. In Alice's case, an organizer reviewed the displayed run and found that it does not contain any obvious errors. In case of errors, explanations are displayed here that give insight into their nature and severity.

*Stdout* Standard output stream (stdout) which was recorded when executing the software. If Alice's software outputs information to stdout, it will be displayed here. However, in the case of runs on test datasets, the amount of information that is displayed can be limited. In the example, the limit is the 100 last chars of the stdout text. This limitation will prevent Alice from outputting problem instances to stdout in order to inspect them. This communication channel can be closed entirely on a per-dataset basis, for example, if confidential data has to be handled.

*Stderr* Standard error output stream (stderr) which was recorded when executing the software. While nothing was recorded in the example, the same filtering is applied as for the stdout stream.

*File List* Directory tree which displays file names and their sizes found in the run. Alice may use this information to determine whether her run has output all the files and directories that are expected, and whether their names and organization are correct.

The run details page will provide Alice with the information necessary to determine whether her remote software execution was successful. Unless the software has been executed on a test dataset, Alice may also download the run for local inspection. If she is satisfied with the run, she may proceed to evaluate it using the evaluation software. The resulting evaluation can again be inspected just like before, whereas the corresponding run details page lists the information pertaining to the evaluation software's run when receiving Alice's software run as input. Figure 7b shows an excerpt of an evaluation run details page that Alice will see. The evaluation software typically prints the evaluation results directly to stdout, however, if the evaluated software run was on a test dataset, the results are blinded by default (i.e., the performance values are replaced by "XXX"). This blinding of the evaluation results upholds blind evaluation. This way, the decision of when, if, and how the evaluation results of a given shared task are released is at the full discretion of its organizers. Moreover, just as with filtering stdout and stderr output, the organizers may adjust blinding on a per-dataset basis.

After completing her evaluation run, Alice is done; she has submitted her software to the virtual machine, made sure it works to the specifications of the shared task by running it on the available datasets and inspecting the runs for errors, and finally executed the evaluation software on her previous software runs. While Alice can now relax, it is time for the organizers of the shared task to get busy.

### 3.5 *Life of an Organizer*

From the perspective of an organizer (say, Bob), using TIRA to manage software submissions for a shared task can be done in three steps: first, configuration of the shared task in TIRA; second, supervision of participant progress; and third, compilation and publication of the task's evaluation results. The interfaces on the right side of Fig. 4 are used for this purpose. The configuration of a shared task is done in a configuration file. Configurable aspects include the datasets and their status (public or hidden), the evaluation software, the command line parameters required for submitted software, and various messages displayed on task-specific web pages. The web interface for task configuration is straightforward; we omit a screenshot for brevity. In terms of supervising his shared task while it is underway, Bob has three further interfaces at his disposal, an overview of participants, an overview of runs of each participant, and the run details of each participant's runs:

*Task Participants (Fig. 8a)* Overview of participants who have configured at least one software for Bob's shared task on their software control page, including their user name, signed in status, numbers of pieces of software that are configured, deleted, and running, and, numbers of runs that are finished, reviewed, and unreviewed. These figures give Bob an idea of whether the participants of his task are actively engaged, but it also hints about problems that may require Bob's attention. The number of deleted pieces of software may indicate that a participant has trouble setting herself up. In the case of Alice, six of seven pieces of software have been deleted, so that it may be the case that Alice had some trouble getting the software configuration right. In the case of Carol, Bob observes that her software has been running for more than six days straight, which may deviate from his expectations. Bob may contact the respective participants and offer his help. Moreover, the number of unreviewed runs indicates that some runs have not yet been checked for errors. To do so, Bob clicks on the review action (the blue eye-icon in the Actions column) to review Alice's runs; he is redirected to the participant's runs page.

*Participant Runs (Fig. 8b)* Overview of a participant's runs on a per-dataset basis, including the software used, run ID, input run, size, numbers of lines, files, and directories, and whether a run has been reviewed. The colorization indicates a run's status with regard to being successful, where red indicates severe errors, yellow indicates warnings, green indicates complete success, and white indicates that the run has not yet been reviewed. Unlike the runs table on Alice's software control

(a)

| 👤 Participants in Shared Task |           |           |         |                 |      |          |            |         |
|-------------------------------|-----------|-----------|---------|-----------------|------|----------|------------|---------|
| User                          | Signed in | Softwares | Deleted | Now Running     | Runs | Reviewed | Unreviewed | Actions |
| Alice                         | yes       | 7         | 6       | none            | 63   | 62       | 1          | 👁️      |
| Carol                         | no        | 1         | 0       | 6 days, 8:37:25 | 4    | 3        | 1          | 👁️      |
| Dan                           | no        | 1         | 0       | none            | 5    | 0        | 5          | 👁️      |
| Eve                           | no        | 3         | 1       | none            | 16   | 16       | 0          | 👁️      |
| Frank                         | no        | 3         | 0       | none            | 56   | 56       | 0          | 👁️      |
| Mallory                       | no        | 1         | 0       | none            | 4    | 0        | 4          | 👁️      |
| Oscar                         | no        | 1         | 0       | none            | 4    | 0        | 4          | 👁️      |
| Peggy                         | no        | 1         | 0       | none            | 4    | 0        | 4          | 👁️      |
| Sybil                         | no        | 3         | 2       | none            | 5    | 5        | 0          | 👁️      |
| Trent                         | no        | 1         | 0       | none            | 4    | 0        | 4          | 👁️      |

(b)

| 📁 Runs of Alice on test-corpus |                                    |                     |       |       |       |      |        |         |
|--------------------------------|------------------------------------|---------------------|-------|-------|-------|------|--------|---------|
| Software                       | Run                                | Input run           | Size  | Lines | Files | Dirs | Review | Actions |
| evaluation                     | 2014-06-22-12-10-00                | 2014-06-22-12-00-00 | 24K   | 36    | 1     | 0    | todo   | 👁️ Ⓜ️   |
| software1                      | 2014-06-22-12-00-00                | none                | 2.2M  | 5180  | 518   | 0    | done   | 👁️ Ⓜ️   |
| software1                      | 2014-06-22-11-00-00                | none                | 2.2M  | 5180  | 518   | 0    | done   | 👁️ Ⓜ️   |
| software1                      | 2014-06-22-10-00-00                | none                | 1.1M  | 2590  | 259   | 0    | done   | 👁️ Ⓜ️   |
| software1                      | 2014-06-22-09-00-00 <sup>DEL</sup> | none                | 0.55M | 1290  | 129   | 0    | done   | 👁️ Ⓜ️   |
| software1                      | 2014-06-22-08-00-00 <sup>DEL</sup> | none                | 1K    | 20    | 2     | 0    | done   | 👁️ Ⓜ️   |

**Fig. 8** TIRA’s web interfaces for organizers to review a task’s participants. (a) Overview of a task’s participants. (b) Overview of a participant’s runs

page, this table shows figures relevant to judging a run’s success to be checked against the expectation for a given dataset: its size and the numbers of lines, files, and directories. Bob has access to all of Alice’s runs including those that have been deleted by Alice (annotated with the superscript “DEL”). Since Bob is task organizer, downloading runs is not restricted. To review the outstanding unreviewed run, Bob clicks on the corresponding review action, redirecting him to the run details page.

*Run Details (Fig. 9)* The run details page corresponds to that which Alice can access. It displays the same information about the run, but there are four differences. (1) it offers a review form in which Bob can enter his review, (2) the standard output streams are not filtered, (3) the output of evaluation software is not blinded, and (4) the button to download the run is always activated. Based on the complete information about the run, Bob can easily review it, which usually takes only a couple of seconds. Bob’s review consists of checking for common errors, such as missing output, extra output, output validity, as well as error messages that have

### Run Details

Overview

|                        |  |
|------------------------|--|
| <b>Software</b>        | evaluation   |
| <b>Run</b>             | 2014-06-22-12-10-00  |
| <b>Input data</b>      | test-data  |
| <b>Input run</b>       | 2014-06-22-12-00-00  |
| <b>Downloadable</b>    | false  |
| <b>Runtime</b>         | 00:00:04 (hh:mm:ss)  |
| <b>Runtime details</b> | 7.04user 14.52system 0:04.10elapsed 52%CPU (0avgtext+0avgdata 85984maxresident)k 0inputs+16outputs (0major+6224minor)pagefaults 0swaps |
| <b>Size</b>            | 24K (15442 bytes)  |
| <b>Lines</b>           | 36   |
| <b>Files</b>           | 2  |
| <b>Directories</b>     | 0  |

Review

This run has not been reviewed, yet.

**Reviewer** Bob

**Errors**

- No errors
- Missing output
- Extra output
- Invalid output
- Error messages in stdout or stderr
- Other kinds of errors; please describe them in the comment below.

**Comment**

**Stdout**

```
python shared-task-evaluation.py -i alice/2014-06-22-12-00-00/output -t test-data -o /tmp/2014-06-22-12-10-00/output/evaluation.txt
```

```
"precision": "0.90081"  
"recall": "0.67283"
```

**Stderr**

**File list**

```
test-data/alice/2014-06-22-12-10-00/output/  
├── [ 246] evaluation.prototext  
└── [ 108] evaluation.txt
```

0 directories, 2 files

Fig. 9 TIRA's web interfaces for organizers to review runs




































































been printed to either standard output stream. These are the common errors that have been observed to occur frequently in previous years (Gollub et al. 2013), whereas Bob has the opportunity to write a short comment about uncommon errors he observes. Bob can supply run verification software for his task that checks runs automatically, however, at least for runs that will be used for the final evaluation results of a shared task, a quick review should be done to foreclose unforeseen errors. This reduces Bob's responsibility for the successful evaluation of Alice's software to a level similar to shared tasks that invite run submissions.

The supervision duties of task organizers cannot be entirely avoided. In shared tasks that invite run submissions, the organizers usually do not have to intervene until after the submission deadline. Only then do they learn how many participants actually submitted a run and how many of the submitted runs are valid as to the specifications of the shared task. In the extreme case, it is only after the run submission deadline, when actual examples of runs on test datasets are available, that the organizers realize that parts of the dataset or the run formats are unfit for their evaluation goals. With software submissions based on TIRA, these risks can be minimized since organizers have a chance to observe early bird participants and make adjustments as the shared task progresses. An added benefit of supervising a shared task using TIRA is that organizers learn early on how many participants actually work toward making a submission to the task, whereas with run submissions, the success or failure of a shared task in terms of number of participants will only become apparent after the run submission deadline. If Bob were to observe that only few participants start using TIRA, he may react by engaging with those who registered but did not start yet, or by advertising the task some more in the community.

Once the submission deadline has passed, and all participants have successfully evaluated their runs on the test datasets of Bob's shared task, he proceeds to reviewing the performance values and to publishing the results. For this purpose, TIRA has an overview of all evaluation runs on a per-dataset basis (see Fig. 10a):

*Evaluations Results Overview* of evaluation runs and the performance results obtained, including user name, software, ID of the evaluation run, ID of the software run that served as input to the evaluation run, and performance values, dependent on the measures computed by a given evaluation software. The colorization of the table cells for both run IDs corresponds to that of the run reviews mentioned above. This helps Bob to decide which are successful evaluations. All evaluation runs of all participants on a given dataset are listed. For example, there are multiple runs for participants Dan and Sybil. Bob gets to decide which of their runs are going to be published; a number of reasonable decision rules are conceivable: (1) all of them (2) the chronologically first or last successful run, (3) the run chosen by the respective participant, or (4) the best-performing run according to a given performance measure. In the Actions column, there are two publishing options, namely publication of evaluation results to the public evaluation results page (the globe icon), and publication of evaluation results to the respective participant (the person icon). As can be seen in the example, Bob has already globally published

(a)

| Evaluations on <i>test-corpus</i> |                          |                     |                                    |           |         |   |
|-----------------------------------|--------------------------|---------------------|------------------------------------|-----------|---------|---|
| User                              | Software                 | Evaluation          | Input run                          | Precision | Recall  | Actions   |
| Alice                             | software1                | 2014-06-22-12-10-00 | 2014-06-22-12-00-00                | 0.90081   | 0.67283 |      |
| Carol                             | software3                | 2014-06-15-17-38-08 | 2014-06-15-17-35-38                | 0.85744   | 0.29661 |      |
| Dan                               | software2 <sup>CEL</sup> | 2014-06-16-17-17-21 | 2014-06-16-16-54-38 <sup>CEL</sup> | 0.96022   | 0.84248 |      |
| Dan                               | software3                | 2014-06-23-20-43-59 | 2014-06-23-20-17-48                | 0.96007   | 0.84511 |      |
| Dan                               | software1                | 2014-06-16-18-03-43 | 2014-06-16-17-21-44                | 0.96243   | 0.83473 |      |
| Eve                               | software1                | 2014-06-01-12-52-02 | 2014-06-21-05-56-23                | 0.82882   | 0.84156 |      |
| Frank                             | software10               | 2014-06-23-13-31-42 | 2014-06-23-13-24-21                | 0.92522   | 0.81819 |      |
| Mallory                           | software1                | 2014-06-20-23-28-21 | 2014-06-17-09-28-40                | 0.87171   | 0.91539 |      |
| Oscar                             | software1                | 2014-06-19-00-54-42 | 2014-06-18-23-50-04                | 0.92757   | 0.88916 |      |
| Peggy                             | software3                | 2014-06-22-03-36-34 | 2014-06-22-03-33-32                | 0.90032   | 0.80267 |      |
| Sybil                             | software2                | 2014-06-22-02-56-09 | 2014-06-22-02-49-41                | 0.90770   | 0.79931 |      |
| Sybil                             | software4                | 2014-06-22-16-55-56 | 2014-06-22-16-49-05                | 0.89179   | 0.80590 |      |
| Trent                             | software5                | 2014-06-15-16-24-05 | 2014-06-15-15-53-28                | 0.86606   | 0.91984 |      |

(b)

| Evaluations on <i>test-corpus</i> |           |         |          |
|-----------------------------------|-----------|---------|----------|
| User                              | Precision | Recall  | Runtime  |
| Alice                             | 0.90081   | 0.67283 | 00:04:17 |
| Carol                             | 0.85744   | 0.29661 | 00:00:56 |
| Dan                               | 0.96007   | 0.84511 | 00:19:32 |
| Eve                               | 0.82882   | 0.84156 | 00:05:18 |
| Frank                             | 0.92522   | 0.81819 | 00:02:49 |
| Mallory                           | 0.87171   | 0.91539 | 00:05:37 |
| Oscar                             | 0.92757   | 0.88916 | 00:57:15 |
| Peggy                             | 0.90032   | 0.80267 | 00:00:31 |
| Trent                             | 0.86606   | 0.91984 | 00:22:10 |

**Fig. 10** TIRA's web interfaces for a task's evaluation results. (a) Overview of a task's evaluation results for organizers. (b) Overview of a task's *published* evaluation results

evaluation runs for all but one participant. Two of Dan's runs are published only to him, and for Sybil's two runs Bob still needs to make a decision.

The published runs appear on a public evaluation results page that can be found on TIRA alongside each shared task. Figure 10b shows the performance values of the evaluations that Bob decided to publish for his shared task. While he proceeds to announce the results to participants as well as to the scientific community, this is not necessarily the end of the story.

Shared task events are organized for a reason, and that reason is not to host an individual run-once competition, but to foster research around a problem of interest. While shared task events are sometimes organized repeatedly, at some point, they are discontinued, whereas later on there are still scientists who want to compare their approach to those of the event's participants. Based on TIRA, this will be easily

possible long after an event is over, since all the evaluation resources required to run an evaluation are hosted and kept in running state. Moreover, if new evaluation datasets appear, all previously developed approaches can be re-evaluated on the new datasets, since they are also kept in running state inside their virtual machines. This way, TIRA paves the way for ongoing, “asynchronous” evaluations around a shared task while ensuring that everyone is evaluated using the exact same experimental setup. That is, of course, as long as TIRA prevails.

## 4 Data Submissions: Crowdsourcing Evaluation Resources

Besides facilitating the submission of software to a given shared task (event), TIRA also opens the door to a more inclusive way of creating an experimental setup around shared tasks: there is no reason why participants should not also submit datasets and alternative performance measures. The organizer of a shared task then grows into the role of a shared task editor, instead of being personally responsible for all but the software. Data submissions for shared tasks have not been systematically studied before, so that no best practices have been established, yet. Asking a shared task event’s participants to submit data is nothing short of crowdsourcing, albeit the task of creating an evaluation resource is by comparison much more complex than average crowdsourcing tasks found in the literature. In what follows, we outline the rationale of data submissions, review important aspects of defining a data submissions task that may inform instructions to be handed out to participants, and detail two methods to evaluate submitted datasets. These are the results from our first-time experience with data submissions to one of our shared task events (Potthast et al. 2015), corresponding to the procedure we followed.

### 4.1 Data Submissions to Shared Tasks: A Rationale

Traditionally, the evaluation resources for a shared task event are created by its organizers—but the question remains: why? The following reasons may apply:

- *Quality control.* The success of a shared task event rests with the quality of its evaluation resources. A poorly built evaluation dataset may invalidate evaluation results, which is one of the risks of organizing shared tasks. This is why organizers have a vested interest in maintaining close control over evaluation resources, and how they are constructed.
- *Seniority.* Senior community members may have the best vantage point in order to create representative evaluation resources.
- *Access to proprietary data.* Having access to an otherwise closed data source (e.g., from a company) gives some community members an advantage over others in creating evaluation resources with a strong connection to the real world.

- *Task inventorship.* The inventor of a new task (i.e., tasks that have not been considered before), is in a unique position to create normative evaluation resources, shaping future evaluations.
- *Being first to the table.* The first one to pick up the opportunity may take the lead in constructing evaluation resources (e.g., when a known task has not been organized as a shared task event before, or, to mitigate a lack of evaluation resources).

All of the above are good reasons for an individual or a small group of scientists to organize a shared task, and, to create corresponding evaluation resources themselves. However, from reviewing dozens of shared task events that have been organized in the human language technologies, none of them are a necessary requirement (Potthast et al. 2014): shared task events are being organized using less-than-optimal datasets, by newcomers to a given research field, without involving special or proprietary data, and without inventing the task in the first place. Hence, we question the traditional connection of shared task event organization and evaluation resource construction. This connection limits the scale and diversity, and therefore the representativeness of the evaluation resources that can be created:

- *Scale.* The number of man-hours that can be invested in the construction of evaluation resources is limited by the number of organizers and their personal commitment. This limits the scale of the evaluation resources. Crowdsourcing may be employed as a means to increase scale in many situations, however, this is mostly not the case when task-specific expertise is required.
- *Diversity.* The combined task-specific capabilities of all organizers may be limited regarding the task's domain. For example, the number of languages spoken by organizers is often fairly small, whereas true representativeness across languages would require evaluation resources from at least all major language families spoken today.

By involving participants in a structured way in the construction of evaluation resources, the organizers may build on their combined expertise, man-power, and diversity. However, there is no free lunch, and outsourcing the construction of evaluation resources introduces the new organizational problem that the datasets created and submitted by third parties must be validated and evaluated for quality.

## 4.2 Inviting Data Submissions

When casting the instructions for data submissions to shared task event, there are a number of desiderata that participants should meet:

- *Data format compliance.* The organizers should agree on a specific data format suitable for the task in question. The format should be defined with the utmost care, since it may be impossible to fix mistakes discovered later on. Experience shows that the format of the evaluation datasets has a major effect on how

participants implement their software for a task. A dataset should comprise a set of problem instances with respect to the task, where each problem instance shall be formatted according to the specifications handed out by the organizers. To ensure compliance, the organizers should prepare a format validation tool, which allows participants to check the format of their to-be-submitted dataset, and whether it complies with the specifications. This way, participants move into the right direction from the start, and less back and forth will be necessary after a dataset has been submitted. The format validation tool should check every aspect of the required data format in order to foreclose any unintended deviation.

- *Annotation validity.* All problem instances of a dataset should comprise ground truth annotations revealing the true solution to the task in question. It goes without saying that all annotations should be valid. Datasets that do not comprise annotations are of course useless for evaluation purposes, whereas annotation validity as well as the quality and representativeness of the problem instances selected by participants determines how useful a submitted dataset will become.
- *Representative size.* The datasets submitted should be of sufficient size, so that dividing them into training and test datasets can be done without sacrificing representativeness, and so that evaluations conducted on the basis of the resulting test datasets are meaningful and not prone to noise.
- *Choice of data source.* The choice of a data source should be left up to participants, and should open the possibility of using manually created data either from the real world or by asking human test subjects to emulate problem instances, as well as automatically generated data based on a computer simulation of problem instances for the task at hand.
- *Copyright and sensitive data.* Participants must ensure that they have the usage rights of the data, for transferring usage rights to the organizers of the shared task event, and for allowing the organizers to transfer usage rights to other participants. The data must further be compliant with privacy laws and ethically innocuous. Dependent on the task at hand and what the organizers of the event desire, accepting confidential or otherwise sensitive data may still be possible by exploiting the datalock: by inviting software submissions to TIRA, the organizers may ensure that the sensitive data does not leak to participants, running submitted software at their site against the submitted datasets.

### 4.3 Evaluating Data Submissions

The construction of new evaluation datasets must be done with the utmost care, since datasets are barely double-checked or questioned again once they have been accepted as authoritative. This presents the organizers who invite data submissions with the new challenge of evaluating submitted datasets, where the evaluation of a dataset should aim at establishing its validity. In general, the organizers of a shared task event that invites data submissions should take care not to advertise submitted

datasets as valid unless they are, since such an endorsement may carry a lot of weight in a shared task's community.

Unlike shared task events that invite algorithmic contributions, the validity of a dataset typically can not be established via an automatically computed performance measure, but requires manual reviewing effort. Though peer-review is one of the traditional means of the scientific community to check and ensure quality, data submissions introduce new obstacles for the following reasons:

- *Dataset size.* Datasets for shared tasks tend to be huge, which renders individual reviewers incapable of reviewing them all. Here, the selection of a statistically representative subset may alleviate the problem, allowing for an estimation of the total amount of errors or other quality issues in a given dataset.
- *Assessment difficulty.* Even if the ground truth of a dataset is revealed, it may not be enough to easily understand and follow the construction principles of a dataset. Additional tools may be required to review problem instances at scale; in some cases, these tools need to solve the task's underlying problem, e.g., to properly visualize problem instances, whereas, without visualization, the review time per problem instance may be prohibitively long.
- *Reviewer bias.* Given a certain assessment difficulty for problem instances, even if the ground truth is revealed, reviewers may be biased to favor easy decisions over difficult ones.
- *Curse of variety.* While shared task events typically address very clear-cut problems, the number of application domains where the task in question occurs may be huge. In these situations, it is unlikely that the reviewers available possess all the required knowledge, abilities, and experience to review and judge a given dataset with confidence.
- *Lack of motivation.* While it is fun and motivating to create a new evaluation resource, reviewing those of others is less so. Reviewers in shared task events that invite data submissions may therefore feel less inclined to invest their time in reviewing other participants' contributions.
- *Privacy concerns.* Some reviewers may feel uncomfortable when passing open judgment on their peers' work for fear of repercussions, especially when they find datasets to be sub-standard. However, an open discussion of the quality of evaluation resources of all kinds is an important prerequisite for progress.

Nevertheless, as long as no third party, impartial reviewers are at hand, as part of their participation, all participants who submit a dataset to a shared task event should be compelled to also peer-review the datasets submitted by other participants. The reviewers may be instructed as follows:

The peer-review is about dataset validity, i.e. the quality and realism of the problem instances. Conducting the peer-review includes:

- *Manual* review of as many examples as possible from all datasets
- Making observations about how the dataset has been constructed
- Making observations about potential quality problems or errors
- Making observations on the realism of each dataset's problem instances

- Writing about your observations in your notebook (make sure to refer to examples from the datasets for your findings).

Handing out the complete submitted datasets for peer-review, however, is out of the question, since this would defeat the purpose of subsequent blind evaluations by revealing the ground truth prematurely. Here, the organizers serve as mediators, splitting submitted datasets into training and test datasets, and handing out only the training portion for peer-review. Obviously, participants who submitted a dataset and who also submitted a software tackling the task in question cannot be subject to blind evaluation on their own dataset, since they possess a copy of the portion of their dataset used as test data. Such conflicts of interest should be highlighted.

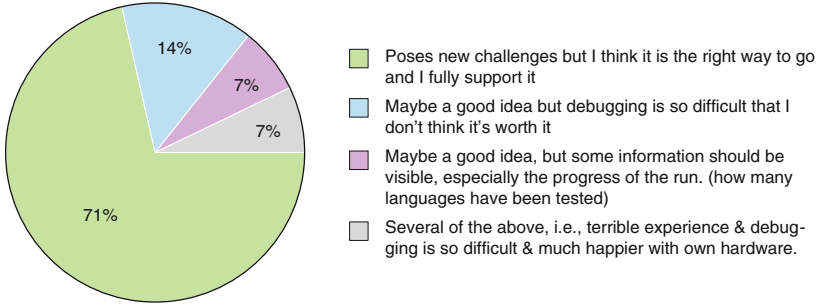
Finally, when a shared task event has previously invited software submissions, this creates ample opportunity to re-evaluate the existing software on the submitted datasets. It becomes possible to evaluate submitted datasets in terms of their difficulty: the performance scores of existing software on submitted datasets, when compared to their respective scores on established datasets, allow for a relative assessment of dataset difficulty. Otherwise, the organizers should set up a baseline software for the task and run that against submitted datasets to allow for a relative comparison among them.

## 5 User Experience and Usage Scenarios

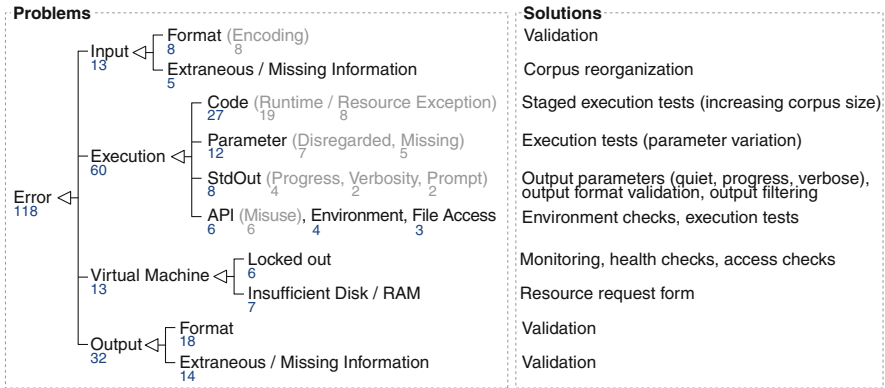
Throughout the years of operating TIRA as a prototypical shared task platform, hosting more than a dozen shared tasks since 2012 until the time of writing, and handling hundreds of software submissions to them, we have gained substantial experience and acquired important insight. To test its limits, we specifically exploited TIRA's capabilities to organize shared task events with advanced experimental setups some of which would not have been feasible otherwise. These usage scenarios and the resulting feedback about user experience inform our ongoing development of TIRA from an already robust prototype to a fully-fledged shared task platform. In what follows, we candidly report successes and failures.

### 5.1 *User Feedback on the Prototype*

We received a great deal of positive user feedback about TIRA from shared task participants as well as their organizers. This has recently been substantiated by the organizers of the CoNLL 2017 shared task, who conducted a user survey; 14 of the 36 participants responded and the results can be seen in Fig. 11. The overwhelming majority of participants, 71%, fully support TIRA and its cause, whereas the remainder ask for improved debugging facilities, significant further improvements, or question the necessity of evaluation platforms altogether. Given



**Fig. 11** Feedback from participants of the CoNLL 2017 shared task



**Fig. 12** Taxonomy of 118 problems that occurred during a busy shared task event that invited software submissions along with technical solutions that identify them automatically. The numbers indicate the amount of errors within each category

the fact that TIRA, in its current form, is an early prototype, and the fact that participants in shared task events such as the one hosted by the CoNLL conference expect high performance of the tools they are supposed to use—after all, research on the task is their primary goal—it is a success that the TIRA prototype achieves this much positive feedback.

To gain more insights into what, precisely, are the errors experienced by TIRA’s users, and how the platform can be further developed to mitigate them, we analyzed the mail correspondence of one of our earlier shared task events (Gollub et al. 2013): 1493 mails were exchanged within 392 conversations, discussing 118 errors. The number of teams experiencing at least one error is 39 from a total of 46, whereas 26 teams experienced at least two errors and one unlucky team 10. The identification of errors and the subsequent discussions induced a significant amount of manual workload. Sometimes, more than one round-trip was necessary to resolve an error. To get a better idea of what kinds of errors occurred and how they can be prevented in the future, we organized them into a taxonomy depicted in; Fig. 12.



In general, input and output errors can be observed in traditional run submission tasks as well, whereas execution errors and virtual machine errors are exclusive to software submission tasks. While the former can be easily identified or prevented by providing format validation and simplifying dataset organization, the latter require more intricate solutions or cannot be identified automatically at all. Since half of all errors are execution errors, the work overhead for organizers is minimized when participants perform execution tests themselves via the web frontend on small-scale trial and training datasets.

## 5.2 *User Retention and Conversion Rate*

The acceptance of shared task platforms also depends on implicit user feedback, namely user retention and conversion rate. The former refers to the retention of users when a shared task event switches from run submissions to software submissions, and the latter to the conversion of registered users to users who submit a software.

*User Retention when Switching to Software Submissions* Given the fact that almost all shared task events today invite run submissions, a barrier to market entry of a shared task platform that solicits software submissions comes in the form of lock-in-effects, where (1) the organizers of successful shared task events are unwilling to switch to another paradigm, since the one at hand works well, and where (2) the organizers of new shared task events follow the example of the majority, who employ run submissions. From those organizers to whom we talked and who considered switching to software submission, they feared that either participants would shy away from the additional overhead, or that organizers would be overwhelmed by it.

Since TIRA was developed within our own shared task series PAN, hosted at the CLEF conference, naturally we employed it there first, starting 2012 with one shared task, and switching all shared tasks to software submission as of 2013. Table 1 shows the numbers of registrations, run/software submissions, and notebook paper submissions for PAN before and after the switch. PAN was growing at the time, and as it turned out, TIRA did not harm the growth of PAN’s shared tasks in terms of participants. Neither of the shared tasks have failed because of TIRA, nor have participation rates dropped compared to the previous “traditional” submission type.

Another success story was the switch of the well-known shared task hosted annually at the CoNLL conference. The CoNLL shared task series has been running since 2000, exchanging the organizer team every 1–2 years. This shared task series is quite renowned and attracts many participants also from well-known institutions. The ACL special interest group of natural language learning (SIGNLL) approached us to inquire about TIRA, and recommended the use of the prototype to implement software submissions for the 2015 edition to its organizers. Ever since, we offer TIRA to the organizers of CoNLL’s shared task events and assist them with the setup. Again, when compared to the participation rates of the many previously

**Table 1** Key figures of the PAN workshop series, hosted at the conferences SEPLN and CLEF

| Statistics    | SEPLN | CLEF |      |      |      |      |      |      |      |
|---------------|-------|------|------|------|------|------|------|------|------|
|               | 2009  | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
| Registrations | 21    | 53   | 52   | 68   | 110  | 103  | 148  | 143  | 191  |
| Runs/software | 14    | 27   | 27   | 48   | 58   | 57   | 54   | 37   | 34   |
| Notebooks     | 11    | 22   | 22   | 34   | 47   | 36   | 52   | 29   | 30   |

Run submission
Software submission

running shared tasks at CoNLL, we did not observe any significant impact after the switch.

*User Retention Within Shared Tasks* Hardly any research has been carried out to date about what makes a shared task (event) successful, or what are success indicators. The number of participants a shared task attracts may be such an indicator, where many consider more participants indicative of more success. This view, however, can easily be refuted, since even a single participant can render a shared task successful, if he or she contributes a groundbreaking approach (i.e., “quality over quantity”). Nevertheless, the number of registrations for a shared task event is indicative of initial interest from the community, whereas the conversion rate from registration to submission is typically significantly less than 100% (as can also be seen in Table 1). Regarding software submissions in general, and TIRA in particular, the question arises whether and how they affect the conversion rate. For lack of data that allows for a statistical analysis, we resort to our experience with organizing the WSDM Cup shared task, which makes for an illuminating showcase in this respect.

The WSDM Cup 2017, organized as part of the ACM WSDM conference, had two tasks for which a total of 140 teams registered, 95 of which ticked the box for participation in the vandalism detection task (multiple selections allowed). This is a rather high number compared to other shared task events. We attribute this success to the facts that the WSDM conference is an A-ranked conference, giving the WSDM Cup a high visibility, that the vandalism detection task was featured on Slashdot,<sup>7</sup> and that we attracted sponsorship from Adobe, which allowed us to award cash prizes to the three winning participants of each task. However, only 35 registered participants actually engaged when being asked for their operating system preferences for their virtual machine on TIRA, 14 of which managed to produce at least one run, whereas the remainder never used their assigned virtual machines at all. In the end, 5 teams made a successful submission to the vandalism detection task by running their software without errors on the test dataset. By contrast, from the 51 registered teams for the second task, triple scoring, 21 successfully submitted their software.

<sup>7</sup><https://developers.slashdot.org/story/16/09/10/1811237>.

Why did so many participants drop out on the vandalism task? We believe that the comparably huge size of the dataset as well as difficulties in setting up their approach on our evaluation platform are part of the reason: each approach had to process gigabytes of data, implement a client-server architecture for this purpose, and it all had to be deployed on a remote virtual machine. The requirement to submit working software, however, may not have been the main cause since the conversion rate of the companion task was much higher. Rather, the combination of dataset size, real-time client-server processing environment, and remote deployment is a likely cause. Note that the vandalism detection task itself demanded this setup, since otherwise it would have been easy to cheat, which is a problem when cash prizes are involved. Finally, the provided baseline systems were already competitive, so that the failure to improve upon them may have caused additional dropouts.

The WSDM Cup taught us an important lesson about the opportunities and limitations of shared task events in general and about evaluation platforms and rule enforcement in particular. On the one hand, competitions like this are important to rally followers for a given task and to create standardized benchmarks. On the other hand, shared task events are constrained to a relatively short period of time and create a competitive environment between teams. I.e., it becomes important to implement a good trade-off in the evaluation setup in order to prevent potential cheating and data leaks, while, at the same time, placing low barriers on the submission procedure. Moreover, there definitely is a trade-off between strict rule enforcement on the one side and scientific progress on the other. For example, only two teams had made a successful submission by the original deadline, while other teams were still struggling with running their approaches. In this case, we erred on the side of scientific progress and accepted submissions up to 8 days past the official deadline, instead of enforcing it. This caused some discussion about the deadline's extensions fairness, where we had to defend the opportunity of more scientific progress over the competitive mindsets of some participants.

Altogether, we do not believe that the low conversion rate of the vandalism detection task was primarily caused by TIRA, but that the task's advanced evaluation setup as a whole did. Otherwise, the companion task would also have had a lower conversion rate. That does not mean that advanced task setups cannot be successful using TIRA. On the contrary, the following section outlines a number of examples where task complexity did not negatively affect conversion rates.

### ***5.3 Advanced Shared Task Setups***

One of the most exciting opportunities that the TIRA prototype offers is to explore new experimental setups of shared tasks. Apart from the aforementioned setup for the WSDM Cup, which involved stream analysis, we have successfully implemented a number of innovative setups that can be seen as testimony to the potential that shared task platforms offer in general. Moreover, they demonstrate the versatility of TIRA in being adapted to different needs.

*Cross-Event Evaluation* One of the primary goals of inviting software submissions in a shared task event is to make re-evaluations of the submitted software on different datasets possible. We grasped at this opportunity immediately after we organized a shared task event for the second time in a row on TIRA, demonstrating this possibility by cross-evaluating software from the previous edition on the then-current evaluation datasets and vice versa. This way, participating in one shared task event corresponds to participating in all of them past, present, and future. Moreover, if a participant submits versions of their software in different years, this makes it possible to track performance improvements. A complete discussion of the results is out of the scope of this section, however, they can be found in Potthast et al. (2013). Cross-year evaluations have become the norm since TIRA is in use for tasks that are organized repeatedly.

*Incorporating External Web Services* When moved into the datalock, TIRA prevents virtual machines from accessing the internet. For some task setups, however, this may be too limiting. For example, an API offered by a third party, which cannot be installed locally inside a virtual machine, may be instrumental to solving a task. The datalock is implemented using a standard firewall, which prevents a virtual machine's virtual network interface from accessing the internet. Naturally, the configuration of the firewall allows for relaxing these access restrictions partially for pre-specified IPs or domains. TIRA implements whitelisting and blacklisting of host names and their associated IP addresses, thereby giving task organizers the liberty to decide whether any, or even just individual participants are allowed to access a given web service.

As part of a series of task events on plagiarism detection, we offered the shared task source retrieval, which is about retrieving a candidate source document from the web for a given suspicious document using a web search engine's API. For sake of realism, we set up a fully-fledged search engine for the ClueWeb corpus, called ChatNoir (Potthast et al. 2012). Given the size of the ClueWeb, we were not able to host more than one instance, let alone one in each participants' virtual machine, so that all submitted software from participants was required to access the search engine. We hence adjusted the datalock firewall setting to give virtual machines access to the host of the web server hosting the ChatNoir's API.

*Shared Task Scale and Resource Allocation* The CoNLL 2017 shared task served as a test of scalability of the evaluation as a service paradigm in general as well as that of TIRA in particular (Zeman et al. 2017). The allocation of an appropriate amount of computing resources (especially CPUs and RAM, whereas disk space is cheap enough) to each participant proved to be difficult, since minimal requirements were unknown. When asked, participants typically request liberal amounts of resources, just to be on the safe side, whereas assigning too much up front would not be economical nor scale well. We hence applied a least commitment strategy with an initial assignment of 1 CPU and 4 GB RAM. More resources were granted on request, the limit being the size of the underlying hardware. When it comes to exploiting available resources, a lot depends on programming prowess, whereas more resources do not necessarily translate into better performance. This is best

exemplified by the fact that with 4 CPUs and 16 GB RAM, the winning team Stanford used only a quarter the amount of resources of the second and third winners, respectively. The team in fourth place was even more frugal, getting by with 1 CPU and 8 GB RAM. All of the aforementioned teams' approaches are within the same ballpark performance, showing that the amount of resources available is no indicator of success at a shared task. Arguably, this may not hold in all conceivable cases.

*Adversarial Shared Tasks* Many tasks that are studied in computer science have an adversarial companion task, where solving one means defeating the solutions for the other. For example, the task of author identification, where given a document, the question to be answered is who wrote it, has its counter in the task of author obfuscation, where given a document, rewrite it so that its author cannot be identified. Here, the performance of an author identifier can only be appreciated if it also defeats all obfuscators available, and vice versa. However, the research community around author identification has never carried out extensive experiments to establish the capabilities of either technology versus the other. This is due to the fact that a serious experiment for a piece of either kind of software requires the procurement of working implementations of *all* existing technologies for the other. Given the amount of author identification technologies that have been proposed in its more than 100 years history, this is an arduous task that can hardly be tackled by any individual or small team of scientists.

The game changes, however, when software submissions come into play. Based on three years worth of software submissions to PAN's author identification tasks (a total of 44 author identifiers representing the state of the art), we have organized for the first time an author obfuscation task, where the goal was to beat each and every one of the identifiers (Potthast et al. 2016). To the best of our knowledge, an evaluation involving adversarial technologies at this scale has hardly any precedent in computer science, if at all, especially when taking into account the low workload at our end. By extension, this demonstrates that, based on a shared task platform like TIRA, compatible or adversarial shared tasks can be composed into pipelines involving even more than just two tasks, thereby opening the door to systematic evaluations of technologies solving very complex tasks.

*Data Submissions* Shared task platforms render data submissions feasible, since the datasets can be immediately evaluated against previously submitted pieces of software for the task in question. One of the longest-running shared tasks at PAN has been text alignment for plagiarism detection. Given that a stable community formed around this task in previous years, and that the data format has not changed throughout the years, we felt confident to experiment with this task and to switch from algorithm development to data submissions. We cast the task to construct an evaluation dataset as follows:

- *Dataset collection.* Gather real-world instances of text reuse or plagiarism, and annotate them.

- *Dataset generation.* Given pairs of documents, generate passages of reused or plagiarized text between them. Apply a means of obfuscation of your choosing.

The task definition has been kept as open as possible, imposing no particular restrictions on the way in which participants approach this task, which languages they consider, or which kinds of plagiarism obfuscation they collect or generate. In particular, the task definition highlights the two possible avenues of dataset construction, namely manual collection, and automatic construction. To ensure compatibility with each other and with previous datasets, however, the format of all submitted datasets had to conform with that of the existing datasets used in previous years. By fixing the dataset format, future editions of the text alignment task may build on the evaluation resources created within the data submission task without further effort, and the pieces of software that have been submitted in previous editions of the text alignment task, available on the TIRA platform, have been re-evaluated on the new datasets. In our case, more than 31 text alignment approaches have been submitted since 2012. To ensure compatibility, we handed out a dataset validation tool that checked all format restrictions. A total of eight datasets have been submitted, offering great variety of languages and data sources. Our approach at validating data submissions for shared tasks followed the procedures outlined in Sect. 4: all participants who submit a dataset have been asked to peer-review the datasets of all other participants, and, all 31 pieces of software that have been submitted to previous editions of our shared task on text alignment were evaluated against the submitted datasets.

We have observed all of the obstacles to peer-review outlined in Sect. 4: some submitted datasets were huge, comprising thousands of generated plagiarism cases; reviewing pairs of entire text documents up to dozens of pages long, and comparing plagiarism cases that may be extremely obfuscated is a laborious task, especially when no tools are around to help; some submitted datasets have been constructed in languages that none of the reviewers speak, except for those who constructed the dataset; and some of the invited reviewers apparently lacked the motivation to actually conduct a review in a useful manner. Nevertheless, the peer-review alongside performance characteristics obtained from the re-evaluation of the 31 plagiarism detectors submitted to the text alignment task in previous years on each submitted dataset gives us confidence that the submitted datasets are reasonably suited to serve as evaluation datasets in the future, significantly increasing the diversity of plagiarism detection corpora available today.

Altogether, as a result of all of these experiments with the design of shared task setups, and because of the reasonable success of the associated shared task events from which ample scientific progress could be extracted, we are happy to say that the TIRA prototype implementation forms a solid foundation to build on in the future. Nevertheless, there are still many avenues of innovation to be explored, and correspondingly many new features need to be developed, let alone the necessary improvement of user experience as well as meeting all requirements outlined at the outset of the chapter.

## 6 Conclusion

The TIRA Integrated Research Architecture offers one of the first and currently the most fully developed platform for shared tasks in the cloud under the evaluation as a service paradigm. Our long-term experience with operating the TIRA prototype, applying it in practice at the PAN workshop at the CLEF conference, at the shared task of the CoNLL conference, and at various other shared task events, has placed TIRA in a unique position to grow, transcending the human language technologies and transferring the concept of shared task evaluations to other branches of computer science. To facilitate this opportunity, TIRA has been released open source, inviting everyone to contribute to its development. This way, we hope to encourage more shared tasks in computer science to gain followership and eventually grow into shared task events. As a platform, TIRA has proven itself, handling more than a dozen shared tasks, some many years in a row, and hundreds of software submissions since its first use in 2012. If successful on a larger scale, TIRA may serve to improve the reproducibility of computer science as a whole.

## References

- Claerbout J, Karrenbach M (1992) Electronic documents give reproducible research a new meaning, pp 601–604. <https://doi.org/10.1190/1.1822162>
- Donoho D, Maleki A, Rahman I, Shahram M, Stodden V (2009) Reproducible research in computational harmonic analysis. *Comput Sci Eng* 11:8–18. <https://doi.org/10.1109/MCSE.2009.15>
- Fokkens A, van Erp M, Postma M, Pedersen T, Vossen P, Freire N (2013) Offspring from reproduction problems: What replication failure teaches us. In: Proceedings of the 51st annual meeting of the association for computational linguistics (Long papers), vol 1. Association for Computational Linguistics, Sofia, pp 1691–1701. <http://www.aclweb.org/anthology/P13-1166>
- Freire J, Fuhr N, Rauber A (2016) Reproducibility of data-oriented experiments in e-science (dagstuhl seminar 16041). *Dagstuhl Rep.* 6(1):108–159. <https://doi.org/10.4230/DagRep.6.1.108>
- Gollub T, Stein B, Burrows S (2012a) Ousting Ivory tower research: towards a web framework for providing experiments as a service. In: Hersh B, Callan J, Maarek Y, Sanderson M (eds) 35th international ACM conference on research and development in information retrieval (SIGIR 2012). ACM, New York, pp 1125–1126. <https://doi.org/10.1145/2348283.2348501>
- Gollub T, Stein B, Burrows S, Hoppe D (2012b) TIRA: configuring, executing, and disseminating information retrieval experiments. In: Tjoa A, Liddle S, Schewe KD, Zhou X (eds) 9th international workshop on text-based information retrieval (TIR 2012) at DEXA. IEEE, Los Alamitos, pp 151–155. <https://doi.org/10.1109/DEXA.2012.55>
- Gollub T, Potthast M, Beyer A, Busse M, Rangel Pardo F, Rosso P, Stamatatos E, Stenno B (2013) Recent trends in digital text forensics and its evaluation—plagiarism detection, author identification, and author profiling. In: Forner P, Müller H, Paredes R, Rosso P, Stein B (eds) Information access evaluation meets multilinguality, multimodality, and visualization. Proceedings of the fourth international conference of the clef initiative (CLEF 2013). Lecture notes in computer science (LNCS), vol 8138, Springer, Heidelberg, pp 282–302

- Hanbury A, Müller H, Balog K, Brodt T, Cormack G, Eggel I, Gollub T, Hopfgartner F, Kalpathy-Cramer J, Kando N, Krithara A, Lin J, Mercer S, Potthast M (2015) Evaluation-as-a-service: overview and outlook. <http://arxiv.org/abs/1512.07454>
- Pedersen T (2008) Empiricism is not a matter of faith. *Comput Linguist* 34(3):465–470 <https://doi.org/10.1162/coli.2008.34.3.465>
- Plesser HE (2018) Reproducibility vs. replicability: a brief history of a confused terminology. *Frontiers in Neuroinformatics* 11:76. <https://doi.org/10.3389/fninf.2017.00076>
- Potthast M, Hagen M, Stein B, Graßegger J, Michel M, Tippmann M, Welsch C (2012) ChatNoir: A search engine for the ClueWeb09 corpus. In: Hersh B, Callan J, Maarek Y, Sanderson M (eds) 35th international ACM conference on research and development in information retrieval (SIGIR 2012). ACM, New York, p 1004. <https://doi.org/10.1145/2348283.2348429>
- Potthast M, Gollub T, Hagen M, Tippmann M, Kiesel J, Rosso P, Stamatatos E, Stein B (2013) Overview of the 5th international competition on plagiarism detection. In: Forner P, Navigli R, Tufis D (eds) Working notes papers of the CLEF 2013 evaluation labs. <http://www.clef-initiative.eu/publication/working-notes>
- Potthast M, Gollub T, Rangel Pardo F, Rosso P, Stamatatos E, Stein B (2014) Improving the reproducibility of PAN's shared tasks: plagiarism detection, author identification, and author profiling. In: Kanoulas E, Lupu M, Clough P, Sanderson M, Hall M, Hanbury A, Toms E (eds) Information access evaluation—multilinguality, multimodality, and interaction. Proceedings of the fifth international conference of the CLEF initiative (CLEF 2014). Lecture notes in computer science (LNCS), vol 8685, Springer, Heidelberg, pp 268–299
- Potthast M, Göring S, Rosso P, Stein B (2015) Towards data submissions for shared tasks: first experiences for the task of text alignment. In: Working notes papers of the CLEF 2015 evaluation labs, CLEF and CEUR-WS.org, CEUR workshop proceedings. <http://www.clef-initiative.eu/publication/working-notes>
- Potthast M, Hagen M, Stein B (2016) Author obfuscation: attacking the state of the art in authorship verification. In: Working notes papers of the CLEF 2016 evaluation labs, CLEF and CEUR-WS.org, CEUR workshop proceedings, vol 1609. <http://ceur-ws.org/Vol-1609/>
- Stodden V (2010) The scientific method in practice: reproducibility in the computational sciences. Tech. Rep. MIT Sloan Research Paper No. 4773-10. <https://doi.org/10.2139/ssrn.1550193>
- Zeman D, Popel M, Straka M, Hajic J, Nivre J, Ginter F, Luotolahti J, Pyysalo S, Petrov S, Potthast M, Tyers F, Badmaeva E, Gokirmak M, Nedoluzhko A, Cinkova S, Hajic jr J, Hlavacova J, Kettnerová V, Uresova Z, Kanerva J, Ojala S, Missilä A, Manning C, Schuster S, Reddy S, Taji D, Habash N, Leung H, de Marneffe MC, Sanguinetti M, Simi M, Kanayama H, de Paiva V, Droганова K, Martínez Alonso H, Çöltekin Ç, Sulubacak U, Uszkoreit H, Macketanz V, Burchardt A, Harris K, Marheinecke K, Rehm G, Kayadelen T, Attia M, Elkahky A, Yu Z, Pitler E, Lertpradit S, Mandl M, Kirchner J, Fernandez Alcalde H, Strnadová J, Banerjee E, Manurung R, Stella A, Shimada A, Kwak S, Mendonca G, Lando T, Nitisaroj R, Li J (2017) CoNLL 2017 shared task: multilingual parsing from raw text to universal dependencies. In: Proceedings of the CoNLL 2017 shared task: multilingual parsing from raw text to universal dependencies. Association for Computational Linguistics, pp 1–19. <https://doi.org/10.18653/v1/K17-3001>



# EaaS: Evaluation-as-a-Service and Experiences from the VISCERAL Project



Henning Müller and Allan Hanbury

**Abstract** The Cranfield paradigm has dominated information retrieval evaluation for almost 50 years. It has had a major impact on the entire domain of information retrieval since the 1960s and, compared with systematic evaluation in other domains, is very well developed and has helped very much to advance the field. This chapter summarizes some of the shortcomings in information analysis evaluation and how recent techniques help to leverage these shortcomings. The term Evaluation-as-a-Service (EaaS) was defined at a workshop that combined several approaches that do not distribute the data but use source code submission, APIs or the cloud to run evaluation campaigns. The outcomes of a white paper and the experiences gained in the VISCERAL project on cloud-based evaluation for medical imaging are explained in this paper. In the conclusions, the next steps for research infrastructures are imagined and the impact that EaaS can have in this context to make research in data science more efficient and effective.

## 1 Introduction

Information retrieval evaluation has largely followed the Cranfield paradigm over the last more than 50 years (Cleverdon et al. 1966; Cleverdon 1962). This means that an information retrieval test collection consisting of documents is created, then topics are defined on the test collection and ground truthing is performed to determine an optimal response. The ground truth can be the relevance of all documents in the collection but is usually only for part of it, done using pooling techniques. The Cranfield tests helped to identify that automatic indexing

---

H. Müller (✉)  
HES–SO Valais, Sierre, Switzerland  
e-mail: [henning.mueller@hevs.ch](mailto:henning.mueller@hevs.ch)

A. Hanbury  
TU Wien, Wien, Austria  
e-mail: [allan.hanbury@tuwien.ac.at](mailto:allan.hanbury@tuwien.ac.at)

of terms had as good or better performance as manually attached keywords of experts by systematic analysis, so their impact on the field of text retrieval was very important. On the basis of this paradigm, information retrieval developed systematic approaches for retrieval system evaluation very early (Jones and van Rijsbergen 1975; Salton 1971) and could thus demonstrate steady scientific progress and also develop commercial applications in many fields by showing the performance obtained over the years. Evaluation campaigns such as TREC<sup>1</sup> (Text REtrieval Conference) (Harman 1992) or CLEF<sup>2</sup> (Cross-Language Evaluation Forum) (Braschler and Peters 2002) developed yearly evaluation cycles for a variety of domains and application scenarios of textual information retrieval. Visual information retrieval systems have been evaluated in ImageCLEF<sup>3</sup> (Image retrieval tasks of CLEF) (Kalpathy-Cramer et al. 2015; Müller et al. 2010) and also the TRECvid campaigns for many years (Smeaton et al. 2003). The impact of these evaluation campaigns both in terms of commercial impact and scholarly impact have also been analyzed for TREC (Rowe et al. 2010), TrecVid (Video retrieval tasks of TREC) (Thornley et al. 2011) and CLEF/ImageCLEF (Tsikrika et al. 2011, 2013). All of these analyses have shown that the commercial impact is massive and that national agencies save much money by supporting such campaigns by sharing resources and allowing larger and more impactful evaluations. The scholarly impact was considered very important, as publicly available test collections foster data reuse and many of the overview papers of the popular evaluation campaigns are highly cited. Participant papers with good results also often get a high citation count as the techniques are often reimplemented or reused (particularly if the source code is also made available). There are also several criticisms of evaluation campaigns, particularly when the tasks are artificial and not related to user needs (Forsyth 2002), or that benchmarking favors small changes to existing algorithms over completely new techniques. In general, challenges improve the performance and not doing any evaluation mainly means that performance improvements cannot be measured. ImageNET (Deng et al. 2009) has also shown that large scale evaluation can lead to disruptive changes, in this case on the use of deep learning for computer vision and object recognition (Krizhevsky et al. 2012).

Evaluation campaigns have many other shortcomings. Even though the topics are usually created with clear user models in mind, often based on surveys or log file analysis (Markonis et al. 2012, 2013; Müller et al. 2007), they only measure static behavior, so changes in user targets or the impact of a user interface and of interaction cannot be measured. For this reason interactive retrieval evaluation was introduced into TREC and CLEF (Borlund and Ingwersen 1997; Gonzalo et al. 2005) to make it possible to measure the human factor in text and image retrieval, which is extremely important for building working systems. Such evaluation is much harder than with static collections but it is very complementary and much can

---

<sup>1</sup><http://trec.nist.gov/>.

<sup>2</sup><http://www.clef-campaign.org/>.

<sup>3</sup><http://www.imageclef.org/>.

be learned. Even with existing test collections and evaluation campaigns, several problems remain and have been reported in the literature. It is often easier to adapt existing data sets to make the challenges easier than to actually improve the techniques (Müller et al. 2002). Even when standard data sets exist, the use of them and the increased performance reported often does not add up (Armstrong et al. 2009b). This is linked to the fact that existing baselines are not well defined and often comparison is not done against the best systems in an evaluation campaign but to less well performing own algorithms, or they are compared against older results, even though better results have been published later on (comparison to a low baseline). Blanco and Zaragoza (2011) shows another problem: when trying many different approaches and manually optimizing them on the data it is possible to get better results but these results are often meaningless, even though they are statistically significant. It is impossible to reproduce such results and statistically speaking significance rules should in this case be multiple and not single hypotheses testing. Often, only positive results are reported and not exactly how these results were obtained. Ioannidis (2005) even goes a step further, stating that most research findings published are false, as small sample sizes are used and often incorrect statistics and strong publication bias towards positive results add to this. Interestingly, the more a domain is competitive the more the results are false, as quick publication and pressure lead to increased incorrectness.

To overcome some of the mentioned problems, several infrastructures have been proposed in the past. In Müller et al. (2001), an online evaluation with a retrieval communication protocol was implemented but it was never used by a large number of systems, even though such a system could guarantee a high level of reproducibility because the ground truth is never released and manual optimizations are difficult. The EvaluatIR initiative (Armstrong et al. 2009a) also developed a framework for evaluation where components could be shared and separately evaluated. Again, such a system could save much time and effort but it was discontinued after only a short period of time. Actually understanding the interplay of the many components of a retrieval system has also been subject of detailed research in the past (Hanbury and Müller 2010). Still, many of the approaches never reached a critical mass and non-integrated systems with intermediate results were often inefficient, and thus not taken up by either researchers or commercial partners. Academic systems to manage evaluation campaigns such as DIRECT have also been used and shown their usefulness (Agosti et al. 2012; Silvello et al. 2017). For good reproducibility both data (Silvello 2018) and code (Niemeyer et al. 2016) need to become citable and reusable easily (Mayernik et al. 2017). Beyond the academic field, machine learning has shown the need for evaluation infrastructures, highlighted by the commercial success of the Kaggle<sup>4</sup> platform and several similar

---

<sup>4</sup><http://www.kaggle.com/>.

initiatives, for example TopCoder<sup>5</sup> and CrowdAI<sup>6</sup> that addresses a more open type of challenges in machine learning.

## 2 Evaluation-as-a-Service

The actual term EaaS<sup>7</sup> (Evaluation-as-a-Service) was detailed in a workshop in early 2015 (Hopfgartner et al. 2015). 12 researchers of several evaluation initiatives and institutions met in Sierre, Switzerland, and discussed their approaches for providing Evaluation-as-a-Service in several environments and with differing approaches. EaaS in this case means that no data sets are distributed but the evaluation part of the campaign is provided as a service. Figure 1 details some of the outcomes of the workshop; also made available in a white paper (Hanbury et al. 2015) in a much more detailed form and with many references to the relevant literature. In Hopfgartner et al. (2018), a shorter version of the main aspects of EaaS was published.

Figure 1 shows the many implications that EaaS has and the various stakeholders in the evaluation environment. Whereas challenge stakeholders can range from challenge organizers and challenge participants (academic and commercial researchers), there are several roles that are often not described in much detail, such as the data providers and human annotators (who might have an important problem that they would like to see solved). In terms of EaaS, infrastructure providers play an important role, for example cloud providers, and also funding agencies that could see important gains in a more efficient global research infrastructure. A similar diversity can be seen in the types of technologies used, the policy aspects of challenges and their infrastructures and also the business parts of it, as all of these can have a strong impact on how data are shared and how well the science advances. EaaS research usually mentions only a few domains of high potential but all these aspects can be taken into account for a full picture. Following the initial workshop, a second workshop was organized in Boston, MA, USA, in November 2015, focusing on the distributed and cloud aspects of the evaluation and with a focus also on medical applications, which is one of the use cases for EaaS with a clearly visible potential (Müller et al. 2016). The various stakeholders from funding agencies, infrastructure and use case providers were invited to this workshop to get a clearer idea of the roles of these partners and their interest.

The example cases that were used as the basis of the white paper include the use of an API for challenges, as in the TREC Microblog task (Ounis et al. 2011) that uses Twitter data that cannot be distributed in another way for copyright reasons.

---

<sup>5</sup><http://www.topcoder.com/>.

<sup>6</sup><http://www.crowdai.org/>.

<sup>7</sup><http://eaas.cc/>.



**Fig. 1** Overview of several aspects and the set of stakeholders in the EaaS field; the figure shows the large number of possibly implied groups and roles (image taken from Hanbury et al. 2015)

The TIRA<sup>8</sup> system uses code submission and then runs the code on the test data in a sandboxed environment on the university servers (Gollub et al. 2012). The CLEF NewsReel task (Hopfgartner et al. 2014) also relies on an API, this time of a news recommendation web page that adds recommendations provided by participating systems to the real recommendations and measures the number of clicks in these provided items to evaluate the quality of the results. VISCERAL (Hanbury et al. 2012) uses virtual machines in the cloud to run a challenge and only a small data set can be seen by the participants. In C-BIBOP<sup>9</sup> (Cloud-Based Image BiOmarker Platform), Docker containers were used to bring the algorithms to the data in a more lightweight way compared to virtual machines. Finally, the BioAsq project is directly included in the process of assigning MeSH terms to new texts and then compares these with the terms that are manually attached to the texts (Tsatsaronis et al. 2015).

<sup>8</sup><http://tira.io/>.

<sup>9</sup><http://cbibop.github.io/>.

Several other challenges have used these concepts as well, for example the Mammography Dream challenge (Trister et al. 2017) that made an unprecedented amount of data available for research in a cloud environment. The Mammography Dream challenge was run in several submission phases that had as objective to improve the initial results. They fostered particularly in the later collaborative phase of the challenge a strong collaboration among the research groups that obtained the best results in the previous phases.

### 3 The VISCERAL Experience

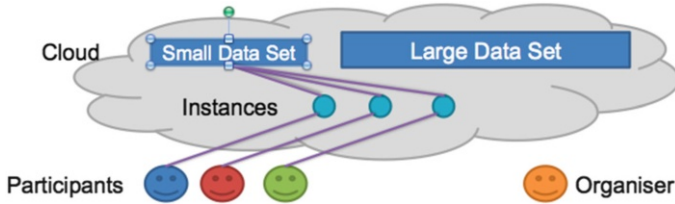
The ideas of the VISCERAL<sup>10</sup> (VISual Concept Extraction challenge in RAdioL-ogy) project were developed on the basis of several challenges and problems of previous evaluation initiatives that the project partners had encountered, notably:

- it is difficult to move *very large data sets*, as even in the Terabyte (TB) range it is currently hard to download data. Sending hard disks through the postal service also becomes cumbersome and prone to physical errors;
- *confidential data* can often not be shared easily but only after manual checking, which becomes infeasible for very large data sets, for example limiting medical data sharing in many cases;
- *quickly changing data sets* cannot be evaluated, as the time to prepare test collections and then transmit them to researchers and obtain results is often already too long and new data have become available in the meantime that need to be taken into account for a final evaluation. This would require a system where algorithms can be run again when new data become available, to always work on the latest data and know what works best.

All these challenges support the idea of moving the algorithms towards the data set rather than the current practice to moving the data to the researchers and their algorithms (Hanbury et al. 2012). The initial idea was to use a cloud infrastructure, in our case the Azure system, to store the entire data set of medical imaging data and only make a small part of it available to researchers directly, keeping the remaining data only accessible to the algorithms and not to the researchers (Langs et al. 2012). Algorithms have actually become much more mobile than the increasingly large data sets. Figure 2 shows the first step of the process, where each participant obtains access to a small data set in the cloud via a virtual machine (VM). This data set makes it possible to get used to the data format and the virtual machine makes it possible to install all necessary tools and test them on the small data set. VMs with both Linux and Windows were available for the participants to avoid creating any limitations. Algorithms and scripts can be tested on the available data, so they will then run automatically on the unknown data.

---

<sup>10</sup><http://www.visceral.eu/>.



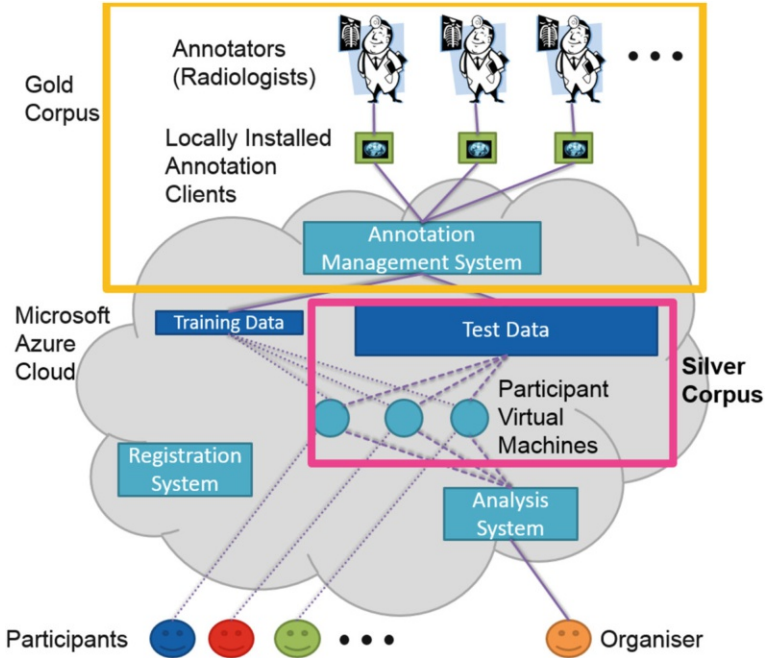
**Fig. 2** The participants each have their own computing instance (VM) in the cloud, linked to a small data set of the same structure as the large one. Software for carrying out the competition objectives is placed in the instances by the participants. The large data set is kept separate (image taken from Hanbury et al. 2012)



**Fig. 3** On the competition deadline, the organiser takes over the instances containing the software written by the participants, links them to the large data set, performs the calculations and evaluates the results (image taken from Hanbury et al. 2012)

The second part of the VISCERAL challenge is shown in Fig. 3. Once the algorithms are ready, then the participant can submit the virtual machine and the proposed algorithms are run on the protected test data. The participants do not have any further access to the machine and only the challenge organizers can access the system to run the algorithms installed on the larger data set. These data thus never get exposed to the researchers and by running the VMs in a sandboxed environment with all communication closed, no data can be communicated out, even if the researchers installed such code on the VMs.

In the case of the VISCERAL benchmarks, the training sets were relatively small and these were made fully available to the participants and the test sets were the large data set (Jimenez-del-Toro et al. 2016). It would also be possible to have training data in the larger data set and thus run training and testing in the cloud directly autonomously. Many more details on the experiences gained in the VISCERAL project can be found in a published book on the project (Hanbury et al. 2017). This book details all challenges that were run (lesion detection, similar case retrieval, organ segmentation), the experiences gained in data annotation and quality controls. The final system to manage the data and challenges can be seen in Fig. 4. The entire data annotation in 3D volumes was also run in the cloud including quality control and several double annotations to measure subjectivity of the tasks (to compare algorithm outcome with the quality of human annotation). A fully automatic evaluation system was developed including the extraction of



**Fig. 4** Final overview of the VISCERAL infrastructure including a system for data annotation and quality control in the cloud and in addition to the cloud-based evaluation (image taken from Hanbury et al. 2015)

many performance measures and an automatically generated leaderboard. When a participant submits a new algorithm all steps are executed automatically and the participant is asked to publish or not the results in the continuous leaderboard of the task.

The system developed also allows many additional possibilities for exploiting the data and the algorithms. Having the algorithms of 15 segmentation tools as executables makes it possible to also run them on new data for which no annotations or ground truth exist. Combining results of several automatic algorithms with label fusion leads to much better results than any single algorithm and we called this outcome a Silver Corpus (Krenn et al. 2016) and showed that it has a very good quality for most of the organs that were segmented in the challenge. This Silver Corpus is now made available to other researchers and can be used to train algorithms such as deep learning algorithms that require large amounts of training data. Thus, it possibly improves results of future systems that can use the additional training data. This also makes it possible to manually annotate only those cases where the algorithms have the highest disagreement, as this can limit the annotation costs with a maximum information gain. Such active learning can massively reduce the efforts of manual annotation and still generate very large and meaningful training data sets.



In general the risk of medical data being exposed to research is not the data itself but rather the risk of matching several data sources or databases, which can possibly allow a re-identification of patients. In the case of VISCERAL no human sees the larger data set, only the algorithms, thus limiting ethical risks to an absolute minimum. Developing such models inside hospitals can even limit the risks further (if security mechanisms for running the code in a protected environment exist). With Docker a light-weight container technology is available that allows to move code and all its dependencies. Hospitals can thus make data of their clinical challenges available for researchers and take advantage of knowledge gained by using the best algorithms. In any case, hospitals will require large computing and storage infrastructures with the advent of digital medicine. Genomics but also the analysis of imaging data and other data sources for decision support will require strong computation in the future, and part of this can also be used for EaaS by sharing data and tasks with the academic world. Even multi-center studies can be envisioned where aggregated data from each client site are combined in a central location. This can be particularly interesting for rare diseases where each institution only encounters few cases over the study period.

## 4 Conclusions

This chapter summarizes the concepts of EaaS and the implications that this can have for scientific challenges in data science, far beyond the initial targets of information retrieval, as in CLEF. Many problems in the academic field such as an exploding number of publications (that are impossible to follow even in a very narrow field (Fraser and Dunstan 2010)) and also the impression that many current publications are not fully correct (Ioannidis 2005) motivate the feeling that new infrastructures for academic research and particularly in data science are necessary. Full reproducibility needs to be created for publications in this field and with executable code containers and digital information this is possible, even in the long term. Storing and linking parameters of all experiments is very important (Di Nunzio and Ferro 2005) to be able to learn a maximum amount from existing experiments. With EaaS, the executable code is available and can be kept in a light-weight manner, via Docker containers, for example. Data are available including ground truth, topics and scripts for running and evaluating tools on the data. If new data become available or if errors in the data are found the code can simply be rerun and all results can be updated.

Besides the credibility of results, it also seems important to make data science more efficient by building fully on the results of other researchers and not reinventing things with minor variations over and over. Good infrastructures can also help with this by facilitating the sharing of code and favoring challenges that allow collaboration between researchers. Working on the same infrastructures and in a similar framework can make sharing code much easier.

Wide availability of all data sets and sharing of recent results on the same data in a simple way can also make it mandatory to compare algorithms with strong baselines, and with the best results obtained on the same data and task at any given time. This would make the quality and possible impact of publications presenting improved results much stronger. It also encourages moving away from focusing purely on quantitative outcomes and rather on interpretation of the potential of techniques and possibly also publishing negative findings, which would help to limit publication bias.

In the current data science environment, most often the groups with the biggest hardware have a strong advantage, as they can optimize parameters of more complex models on more training data and often obtain better results. By running challenges in a central infrastructure all participants have access to the same computational power, so this disadvantage would vanish and it would even make it possible to compare algorithm effectiveness and efficiency, as usually a tradeoff between the two is the main objective. With many research centers now using virtual machines and virtualized environments, it does not seem to matter too much anymore where the physical servers are, as long as they remain accessible and if quick access to the data is possible. In such an environment, EaaS is a very natural choice. Using Docker containers instead of VMs is in our experience a big advantage, as the installation overhead is very low and portability is much higher than with VMs.

In the future we expect evaluation campaigns such as CLEF to have access to their own research infrastructures and to make them available for participating researchers. The question of who bears the costs for data storage and computation need to be answered. Overall, the costs will be lower, so funding bodies should have a strong interest in having such a framework installed for scientific research. On a European level, the EOSC (European Open Science Cloud) is a candidate to supply storage and computation for such an approach in data science challenges. Likely there are still barriers to this approach but it is a question of time before models similar to EaaS will become the most common form of performing data science.

**Acknowledgements** The work leading to the chapter was partly funded by the EU FP7 program in the VISCERAL project and the ESF via the ELIAS project. We also thank all the participants of the related workshops for their input and the rich discussions.

## References

- Agosti M, Di Buccio E, Ferro N, Masiero I, Peruzzo S, Silvello G (2012) Directions: design and specification of an ir evaluation infrastructure. In: Springer (ed) Multilingual and multimodal information access evaluation—third international conference of the cross-language evaluation forum, LNCS, vol 7488, pp 88–99
- Armstrong TG, Moffat A, Webber W, Zobel J (2009a) EvaluatIR: an online tool for evaluating and comparing ir systems. In: Proceedings of the 32nd international ACM SIGIR conference, SIGIR'09. ACM, New York, p 833. <http://doi.acm.org/10.1145/1571941.1572153>

- Armstrong TG, Moffat A, Webber W, Zobel J (2009b) Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceeding of the 18th ACM conference on Information and knowledge management, CIKM'09*. ACM, New York, pp 601–610. <http://doi.acm.org/10.1145/1645953.1646031>
- Blanco R, Zaragoza H (2011) Beware of relatively large but meaningless improvements. Tech. rep., Yahoo Research
- Borlund P, Ingwersen P (1997) The development of a method for the evaluation of interactive information retrieval systems. *J Doc* 53:225–250
- Braschler M, Peters C (2002) The CLEF campaigns: evaluation of cross-language information retrieval systems. *CEPIS UPGRADE III* 3:78–81
- Cleverdon CW (1962) Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Tech. rep., Aslib Cranfield Research Project, Cranfield
- Cleverdon C, Mills J, Keen M (1966) Factors determining the performance of indexing systems. Tech. rep., ASLIB Cranfield Research Project, Cranfield
- Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: *IEEE conference on computer vision and pattern recognition, CVPR 2009*, pp 248–255
- Di Nunzio GM, Ferro N (2005) Direct: a system for evaluating information access components of digital libraries. In: *International conference on theory and practice of digital libraries*. Springer, Berlin, pp 483–484
- Forsyth DA (2002) Benchmarks for storage and retrieval in multimedia databases. In: *SPIE Proceedings of storage and retrieval for media databases*, vol 4676, San Jose, pp 240–247 (SPIE photonics west conference)
- Fraser AG, Dunstan FD (2010) On the impossibility of being expert. *BMJ* 341:c6815
- Gollub T, Stein B, Burrows S (2012) Ousting ivory tower research: towards a web framework for providing experiments as a service. In: *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval*. ACM, New York, pp 1125–1126
- Gonzalo J, Clough P, Vallin A (2005) Overview of the CLEF 2005 interactive track. In: *Working notes of the 2005 CLEF workshop*, Vienna
- Hanbury A, Müller H (2010) Automated component-level evaluation: present and future. In: *International conference of the cross-language evaluation forum (CLEF)*. Lecture notes in computer science (LNCS), vol 6360. Springer, Berlin, pp 124–135
- Hanbury A, Müller H, Langs G, Weber MA, Menze BH, Fernandez TS (2012) Bringing the algorithms to the data: cloud-based benchmarking for medical image analysis. In: *CLEF conference*. Lecture notes in computer science. Springer, Berlin
- Hanbury A, Müller H, Balog K, Brodt T, Cormack GV, Eggel I, Gollub T, Hopfgartner F, Kalpathy-Cramer J, Kando N, Krithara A, Lin J, Mercer S, Potthast M (2015) Evaluation-as-a-service: overview and outlook. *arXiv* 1512.07454
- Hanbury A, Müller H, Langs G (eds) (2017) *Cloud-based benchmarking of medical image analysis*. Springer, Berlin
- Harman D (1992) Overview of the first text REtrieval conference (TREC-1). In: *Proceedings of the first text REtrieval conference (TREC-1)*, Washington, pp 1–20
- Hopfgartner F, Kille B, Lommatzsch A, Plumbaum T, Brodt T, Heintz T (2014) Benchmarking news recommendations in a living lab. In: *International conference of the cross-language evaluation forum for European languages*. Springer, Berlin, pp 250–267
- Hopfgartner F, Hanbury A, Müller H, Kando N, Mercer S, Kalpathy-Cramer J, Potthast M, Gollub T, Krithara A, Lin J, Balog K, Eggel I (2015) Report on the evaluation-as-a-service (EAAS) expert workshop. *ACM SIGIR Forum* 49(1):57–65
- Hopfgartner F, Hanbury A, Müller H, Balog K, Brodt T, Cormack GV, Eggel I, Gollub T, Kalpathy-Cramer J, Kando N, Krithara A, Lin J, Mercer S, Potthast M (2018) Evaluation-as-a-service in the computational sciences: overview and outlook. *J Data Inf Qual* 10(4):15
- Ioannidis JP (2005) Why most published research findings are false. *PLoS Med* 2(8):e124

- Jimenez-del-Toro O, Müller H, Krenn M, Gruenberg K, Taha AA, Winterstein M, Eggel I, Foncubierta-Rodríguez A, Goksel O, Jakab A, Kontokotsios G, Langs G, Menze B, Salas Fernandez T, Schaer R, Walleyo A, Weber MA, Dicente Cid Y, Gass T, Heinrich M, Jia F, Kahl F, Kechichian R, Mai D, Spanier AB, Vincent G, Wang C, Wyeth D, Hanbury A (2016) Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL anatomy benchmarks. *IEEE Trans Med Imaging* 35(11):2459–2475
- Jones KS, van Rijsbergen C (1975) Report on the need for and provision of an ideal information retrieval test collection. *British Library Research and Development Report 5266*, Computer Laboratory, University of Cambridge
- Kalpathy-Cramer J, García Seco de Herrera A, Demner-Fushman D, Antani S, Bedrick S, Müller H (2015) Evaluating performance of biomedical image retrieval systems: overview of the medical image retrieval task at ImageCLEF 2004–2014. *Comput Med Imaging Graph* 39:55–61
- Krenn M, Dorfer M, Jimenez-del-Toro O, Müller H, Menze B, Weber MA, Hanbury A, Langs G (2016) Creating a large-scale silver corpus from multiple algorithmic segmentations. In: Menze B, Langs G, Montillo A, Kelm M, Müller H, Zhang S, Cai W, Metaxas D (eds) *Medical computer vision: algorithms for big data: international workshop, MCV 2015, Held in Conjunction with MICCAI 2015, Munich, Germany, October 9, 2015, Revised Selected Papers*, Springer International Publishing, pp 103–115
- Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) *Advances in neural information processing systems*, vol 25, Curran Associates, pp 1097–1105
- Langs G, Hanbury A, Menze B, Müller H (2012) VISCERAL: towards large data in medical imaging—challenges and directions. In: Greenspan H, Müller H, Syeda-Mahmood T (eds) *Medical content-based retrieval for clinical decision support. Lecture notes in computer science*. Springer, Berlin, pp 92–98
- Markonis D, Holzer M, Dungs S, Vargas A, Langs G, Kriewel S, Müller H (2012) A survey on visual information search behavior and requirements of radiologists. *Methods Inf Med* 51(6):539–548
- Markonis D, Baroz F, Ruiz de Castaneda RL, Boyer C, Müller H (2013) User tests for assessing a medical image retrieval system: a pilot study. *Stud Health Technol Inform* 192:224–228
- Mayernik MS, Hart DL, Mauli DL, Weber NM (2017) Assessing and tracing the outcomes and impact of research infrastructures. *J Am Soc Inf Sci Technol* 68(6):1341–1359
- Müller H, Müller W, Marchand-Maillet S, Squire DM, Pun T (2001) Automated benchmarking in content-based image retrieval. In: *Proceedings of the second international conference on multimedia and exposition (ICME'2001)*. IEEE Computer Society, Silver Spring, pp 321–324
- Müller H, Marchand-Maillet S, Pun T (2002) The truth about corel-evaluation in image retrieval. In: Lew MS, Sebe N, Eakins JP (eds) *Proceedings of the international conference on the challenge of image and video retrieval (CIVR 2002)*. *Lecture notes in computer science (LNCS)*, vol 2383. Springer, Berlin, pp 38–49
- Müller H, Boyer C, Gaudinat A, Hersh W, Geissbuhler A (2007) Analyzing web log files of the Health On the Net HONmedia search engine to define typical image search tasks for image retrieval evaluation. *Stud Health Technol Inform* 129(Pt 2):1319–1323
- Müller H, Clough P, Deselaers T, Caputo B (eds) (2010) *ImageCLEF—experimental evaluation in visual information retrieval*. In: *The Springer international series on information retrieval*, vol 32. Springer, Berlin
- Müller H, Kalpathy-Cramer J, Hanbury A, Farahani K, Sergeev R, Paik JH, Klein A, Criminisi A, Trister A, Norman T, Kennedy D, Srinivasa G, Mamonov A, Preuss N (2016) Report on the cloud-based evaluation approaches workshop 2015. *ACM SIGIR Forum* 51(1):35–41
- Niemeyer KE, Smith AM, Katz DS (2016) The challenge and promise of software citation for credit, identification, discovery, and reuse. *J Data Inform Qual* 7(6):161–165
- Ounis I, Macdonald C, Lin J, Soboroff I (2011) Overview of the trec-2011 microblog track. In: *Proceedings of the 20th text REtrieval conference (TREC 2011)*, vol 32

- Rowe BR, Wood DW, Link AN, Simoni DA (2010) Economic impact assessment of NIST text retrieval conference (TREC) program. Technical report project number 0211875, National Institute of Standards and Technology
- Salton G (1971) The SMART retrieval system, experiments in automatic document processing. Prentice Hall, Englewood Cliffs
- Silvello G (2018) Theory and practice of data citation. *J Assoc Inf Sci Technol* 69:6–20
- Silvello G, Bordea G, Ferro N, Buitelaar P, Bogers T (2017) Semantic representation and enrichment of information retrieval experimental data. *Int J Digit Libr* 18(2):145–172
- Smeaton AF, Kraaij W, Over P (2003) TRECVID 2003: an overview. In: *Proceedings of the TRECVID 2003 conference*
- Thornley CV, Johnson AC, Smeaton AF, Lee H (2011) The scholarly impact of TRECVID (2003–2009). *J Am Soc Inf Sci Technol* 62(4):613–627
- Trister AD, Buist DS, Lee CI (2017) Will machine learning tip the balance in breast cancer screening? *JAMA Oncol* 3(11):1463–1464
- Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, Weissenborn D, Krithara A, Petridis S, Polychronopoulos D, et al (2015) An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinf* 16(1):138
- Tsikrika T, García Seco de Herrera A, Müller H (2011) Assessing the scholarly impact of ImageCLEF. In: *CLEF 2011. Lecture notes in computer science (LNCS)*. Springer, Berlin, pp 95–106
- Tsikrika T, Larsen B, Müller H, Endrullis S, Rahm E (2013) The scholarly impact of CLEF (2000–2009). In: *information access evaluation, multilinguality, multimodality, and visualization*. Springer, Berlin, pp 1–12

**Part III**  
**Multilingual and Multimedia Information**  
**Retrieval**

# Lessons Learnt from Experiments on the Ad Hoc Multilingual Test Collections at CLEF



Jacques Savoy and Martin Braschler

**Abstract** This chapter describes the lessons learnt from the ad hoc track at CLEF in the years 2000 to 2009. This contribution focuses on Information Retrieval (IR) for languages other than English (monolingual IR), as well as bilingual IR (also termed “cross-lingual”; the request is written in one language and the searched collection in another), and multilingual IR (the information items are written in many different languages). During these years the ad hoc track has used mainly newspaper test collections, covering more than 15 languages. The authors themselves have designed, implemented and evaluated IR tools for all these languages during those CLEF campaigns. Based on our own experience and the lessons reported by other participants in these years, we are able to describe the most important challenges when designing a IR system for a new language. When dealing with bilingual IR, our experiments indicate that the critical point is the translation process. However, currently online translating systems tend to offer rather effective translation from one language to another, especially when one of these languages is English. In order to solve the multilingual IR question, different IR architectures are possible. For the simplest approach based on query translation of individual language pairs, the crucial component is the merging of the intermediate bilingual results. When considering both document and query translation, the complexity of the whole system represents clearly a main issue.

---

J. Savoy (✉)

Computer Science Department, University of Neuchâtel, Neuchâtel, Switzerland

e-mail: [Jacques.Savoy@unine.ch](mailto:Jacques.Savoy@unine.ch)

M. Braschler

Institut für Angewandte Informationstechnologie, Zürich University of Applied Sciences ZHAW,  
Winterthur, Switzerland

e-mail: [bram@zhaw.ch](mailto:bram@zhaw.ch)

© Springer Nature Switzerland AG 2019

N. Ferro, C. Peters (eds.), *Information Retrieval Evaluation*

*in a Changing World*, The Information Retrieval Series 41,

[https://doi.org/10.1007/978-3-030-22948-1\\_7](https://doi.org/10.1007/978-3-030-22948-1_7)

## 1 Introduction

In the field of natural language research and applications, the English language is getting the most attention. With the growing presence of web sites not written in English, there is an increasing demand for effective tools to manipulate content in other natural languages. Such interest has also been supported by the process of globalization. Looking at the world around us, one can see many documents in digital libraries, newspapers, government archives and records, as well as legal and court decision documentation not written in English. For example, the European Union counts 24 official languages in 2017, and for each of them, effective IR tools must be designed, implemented, and evaluated. This objective corresponds to one of the main purposes of the mono-, bi-, and multilingual ad-hoc tracks in the CLEF evaluation campaigns.

At a first glance, one can think that a simple adaptation of approaches for handling English should be enough. After all, a cursory observer may assume that all European languages belong to the same Indo-European language family and stem from the same source. This assumption is not true. First, the Finnish, Hungarian and Estonian languages are members of the Uralic family, while the Maltese language is related to the Semitic group. All these languages serve as official EU languages. Second, morphology and word construction vary considerably between members of the Indo-European family, reducing the effectiveness of a simple adaptation from English. As a possible language-independent solution, one can design and implement search models based on the character  $n$ -grams approach (McNamee et al. 2009). Such a text representation approach was also shown effective for Chinese, Japanese or Korean languages (Savoy 2005). To reflect the language differences more closely, the current chapter describes an overview of approaches taking into account the morphology differences between the different languages. Most experiments in the CLEF ad hoc track have followed this approach.

Some of the use cases associated with accessing sources written in languages other than English and more generally in a multilingual context are as follows: in multilingual countries such as Switzerland, institutions such as the Federal Supreme Court may have to document legal cases, or parts of them, in one of the national languages (German, French, or Italian), depending on the involved parties, without providing translations into the other official languages. The information contained in these documents is still relevant for the whole country regardless of the language chosen. Also worth considering are the books and documents available in various languages in our libraries, in multinational companies or large international organizations (e.g., World Trade Organization, European Parliament or United Nations), where the typical user needs to overcome various language barriers. For example, users may write a request in one language and yet wish to retrieve documents written in one or more other languages. Frequently, users have some degree of proficiency in other languages that allows them to read documents, but not to formulate a query in that language or, at least, not to provide reliable search terms to retrieve the documents being searched. In other circumstances,



monolingual users may want to retrieve documents in another language and then automatically or manually translate the texts retrieved into their own language. Finally, there are many documents in other languages containing information in non-textual formats such as images, graphics, and statistics that could be made accessible to monolingual users, based on requests written in a different language.

Based on more than a decade of experiments on developing CLIR systems, the rest of this paper is organized as follows. The next section describes the main problems when designing and implementing an IR system for a new language (monolingual IR). Section 3 discusses briefly the various solutions that can be applied to develop a bilingual system that does “cross-lingual” IR, returning information items in a language other than that used for the request. The description of different multilingual IR architectures is presented in Sect. 4 together with their advantages and drawbacks. Our main findings are summarized in the conclusion.

## 2 Monolingual (Non-English) Information Retrieval

The implementation of IR systems is conceptually subdivided into two major phases: the indexing and the matching phases. When moving from the English language to more (potentially all) languages, we have to re-think both phases. We start our discussion with the indexing phase, which is often implemented in the form of an indexing pipeline, where information items (and the requests) are methodically transformed into representations suitable for matching. Usually, the first step is to extract the words (tokenization). As white space characters and punctuation symbols are used to denote the word boundaries for all European languages, the tokenization to be applied for these languages does not differ fundamentally from that used for English (note, however, minor differences such as the handling of contractions, like “aujourd’hui” (today) in French). After being able to determine words, the morphology of the underlying language is of prime importance. Thus, knowing the part-of-speech (POS) of very frequent words is useful to define an appropriate stopword list as indicated in Sect. 2.1. Moreover, the word formation construction varies from one language to another. Thus, there is a real need to create an effective stemmer for each language as shown in Sect. 2.2. Finally, in Sect. 2.3 we explore the matching phase. Findings indicate that fundamental concepts used in IR weighting schemes such as term frequency ( $tf$ ), inverse document frequency ( $idf$ ), and length normalization are valid across all languages.

### 2.1 Stopword List

Information retrieval weighting schemes suffer from a drop in effectiveness if extremely frequent non-content bearing words are present. In such cases, the  $idf$ -weight that should account for global frequency of terms no longer balances

the contribution to the overall score. Typically, function words (determiners, prepositions, conjunctions, pronouns, and auxiliary verbal forms) are affected. Assuming that these do not convey important meaning, they can be regrouped in a stopword list to be ignored during the indexing procedure. For all languages, the identification of determiners, prepositions (or, for some languages, postpositions), conjunctions, and pronouns does not present a real difficulty. Delimiting precisely whether an auxiliary verb form must appear or not in a stopword list is less clear. Forms such as those related to the verb “to be” and “to have” are good candidates for inclusion. For the modal verbs (e.g., can, would, should), the decision is debatable. For example, one can decide that “shall” must be included but not “can” or “must”.

Reflecting the root cause of the problem (the very high occurrence count of some of these words), we can opt for a frequentist perspective instead of using POS information. In this case, a stopword list can be defined as the  $k$  most frequent words (with  $k = 10$  to 500) in a given corpus or language (Fox 1990). With this strategy, some recurrent lexical words of the underlying corpus will appear in the top  $k$  most frequent words. For example with newspaper collections, very frequent words (e.g., government, president, world), names (e.g., France, Obama) or acronyms (e.g., PM, UK, GOP) will also appear in the top of the resulting ranked list. This would seem undesirable, but note that words that appear with very high frequency are in any case badly suited to discriminate documents even should they be content-bearing.

After applying one of the two previous solutions, an inspection phase must verify whether the presence of a word in a stopword list could be problematic such as, for example, with homographs (e.g., “US” can be a country or a pronoun). For example, in French the word “or” can be translated into “thus/now,” or “gold” while the French word “est” can correspond to “is” or “East”. This verification must not be limited to the vocabulary but must take into account some acronyms (e.g., the pronoun “who” must be separated from the acronym “WHO” (World Health Organization) due, in this case, to the fact that uppercase letters are replaced by the lowercase equivalents.

Applying a stopword list generally improves the overall mean average precision (MAP). The precise value of such improvement depends on the language and the IR model, but an relative average enhancement may vary from 11.7% (English) to 17.4% (French). However, with either a long or a rather short stopword list, the retrieval effectiveness tends to be similar (MAP difference around 1.6% for the English language, 1.2% for French (Dolamic and Savoy 2010c)).

Some commercial IR systems consider that functional words may be entered by the user (e.g., search engines on the Web) or that they can be useful to specify the meaning more closely (e.g., specialized IR systems with “vitamin A”). Therefore, the size of the stopword list can be limited to a few very frequent words. As an extreme case and for the English language, the stopword list could be limited to a single entry (the article “the”) (Moulinier 2004). Since stopword elimination always implies an information loss, however small, one is advised to use robust weighting schemes that allow the use of short stopword lists (Dolamic and Savoy 2010c).

## 2.2 *Morphological Variations*

A first visual difference between an English text and a document written in another European language could be the presence of a non-Latin script such as, for example, when the Cyrillic alphabet is employed for the Russian and Bulgarian languages. Another visual distinction is often the presence of diacritics (e.g., “élite”, “Äpfel” (apples), “leão” (lion)). Different linguistic functions are attached to those additional glyphs such as discriminating between singular (“Apfel”) and plural form (“Äpfel”), between two possible meanings (e.g., “tâche” (task) or “tache” (mark, spot)), or specifying the pronunciation. Keeping those diacritics or replacing them with the corresponding single letter modifies marginally the mean average precision (MAP), usually not in a significant way, and not always in the same direction. Note also that in some languages, it may be permissible to skip the writing of diacritics in certain circumstances, which may lead to an uneven use throughout a textual corpus. In such cases, elimination of diacritics may be advisable (e.g., in French, diacritics are usually not written in upper-case text). In German, umlauts are replaced if the corresponding keys are not available on a keyboard (e.g., “Zürich” can be written “Zuerich”).

To achieve an effective semantic matching between words appearing in the user’s request and the document surrogates, the indexing procedure must ignore small variations between a word stem (e.g., friend) and the various surface forms (e.g., friends). Such morphological variations may for example reflect the word’s function in a sentence (grammatical cases), the gender (masculine, feminine, neutral), and the number (singular, dual, plural). For verbs, the tense, the person, and the mode may generate additional variations. These morphological variations are marked by inflectional suffixes that must be removed to discover the word stem. Of course, one can always find some exceptions such as, for example, having a plural form not always related to the singular one (e.g., “aids”, the syndrome, and “aid” for help) while some words usually appear in only one form (e.g., scissors).

The English language has a comparatively simple inflectional morphology. For example, the noun plural form is usually indicated by the “-s” suffix. To denote the plural form in Italian, the last vowel (usually “-o”, “-e”, or “-a” for masculine nouns, “-a” or “-e” in feminine) must be changed into “-i” or “-e”. In German, the plural can be indicated by a number of suffixes or transformations (e.g., “Apfel” into “Äpfel” (apple), “Auge” into “Augen” (eye), “Bett” into “Betten” (bed)). Variations in grammatical cases (nominative, accusative, dative, etc.) may imply the presence of a suffix (as, for example, the “’s” in “Paul’s book”). In German, the four grammatical cases and three genders may modify the ending of adjectives or nouns. The same is valid for other languages such as Russian (6 cases), Czech (7 cases), Finnish (15 cases) or Hungarian (17 cases). As a simple indicator to define the morphological complexity of a language, one can multiply the number of possible genders, numbers, and grammatical cases. With this measure, the Italian or French language has a complexity of  $2$  (genders)  $\times$   $2$  (numbers) =  $4$  (no grammatical case denoted by a suffix) while the German complexity is  $3 \times 2 \times 4 = 24$ .

New words can also be generated by adding derivational affixes. In IR, we assume that adding a prefix will change the meaning (e.g., bicycle, disbelief) and thus only suffix removal is usually considered (e.g. friendly, friendship).

Based on our experiments, it is not always clear whether a light stemmer (removing only inflectional suffixes or part of them) or an aggressive stemmer removing both inflectional and derivational suffixes proposes the best solution. For the English language, the conservative S-stemmer (Harman 1991) removes only the plural suffix while Porter's stemmer (Porter 1980) is a more aggressive approach. Such algorithmic or rule-based stemmers ignore word meanings and tend to make errors, usually due to over-stemming (e.g., "organization" is reduced to "organ") or to under-stemming (e.g., "European" and "Europe" do not conflate to the same root). In both cases, we suggest concentrating mostly on nouns and adjectives, and ignoring most of the verbal suffixes. Usually the meaning of a sentence can be determined more precisely when focusing more on the noun phrases than on the verbs.

While stemming approaches are normally designed to work with general texts, a stemmer may also be specifically designed for a given domain (e.g., medicine) or a given document collection, such as that developed by Paik and Parai (2011) or Paik et al. (2013) which used a corpus-based approach. This stemming approach reflects the language usage more closely (including word frequencies and other co-occurrence statistics), instead of a set of morphological rules.

A study demonstrates however that using a morphological analysis both light or more aggressive stemmers tend to produce statistically similar performance for the English language (Fautsch and Savoy 2009). When the stemmed words are shown to the user, we suggest applying a light stemmer approach for which the relationship between the surface form and the transformed one is usually simple and more understandable.

Using the CLEF test collections and the Okapi IR model (Robertson et al. 2000), one can find the following retrieval improvement (MAP) with a light stemmer over a non-stemming approach: +7% with the English language (Fautsch and Savoy 2009), +11% for German (Savoy 2006), +28% for Portuguese (Savoy 2006), +34% for French (Savoy 2006), +38% for Bulgarian (Savoy 2008a), +44% for Czech (Dolamic and Savoy 2009b), +55% for Hungarian (Savoy 2008b), and +96% with the Russian language (Dolamic and Savoy 2009a). Working with a morphologically rich language presenting numerous inflectional suffixes (e.g., Hungarian (Savoy 2008b)), even for names (e.g., Czech (Dolamic and Savoy 2009b); Russian (Dolamic and Savoy 2009a)), the presence of a stemming procedure is mandatory to achieve good retrieval effectiveness. Such IR tools are freely available for many languages.<sup>1</sup>

The choice between a light or a more aggressive suffix-stripping procedure for many languages remains not completely obvious. When looking only at the mean performance difference between a light and an aggressive stemmer, the variation

---

<sup>1</sup>Freely available at [www.unine.ch/info/clef/](http://www.unine.ch/info/clef/) or at [tartarus.org/martin/PorterStemmer/](http://tartarus.org/martin/PorterStemmer/).

depends on the language, IR model, and test collection. For the English language, the average performance differences between a light (S-stemmer) and Porter's stemmer is 1% over five IR models and in favor of Porter's solution. This difference is however not statistically significant. With the Russian language, the difference is also 1% in average, but in favor of a light approach. For French, the aggressive stemmer performs, in mean, 6% better, but only 3% for Czech. Thus no clear and definitive conclusion can be reached when comparing the effectiveness of a light vs. more aggressive stemmer.

Finally, compounding, i.e. a word formation process where new words are formed based on multiple simpler "components" (e.g., ghostwriter, dishwasher), is another linguistic construction that can affect the IR quality. This form is active in many languages (e.g., "capogiro" (dizziness) in Italian, "rakkauskirje" (love (rakkaus) and letter (kirje) in Finnish) but especially in German compounding is frequent and raises a specific challenge (Sanders 2010). First, this language allows long constructions (e.g., "Lebensversicherungsgesellschaftsangestellter" = "Leben" + s + "Versicherung" + s + "Gesellschaft" + s + "Angestellter" for life + insurance + company + employee)). Second, the same concept can equally be expressed using a compound term (e.g., "Computersicherheit") or a noun-phrase ("die Sicherheit für Computer"). As one form can appear in a relevant text and the second in the query, this aspect renders semantic matching more difficult. Thus, for the German language, a decompounding procedure (Chen 2004; Savoy 2003b) must be applied to achieve higher effectiveness. Such an automatic word decomposition can improve the MAP by 23% for short queries (title-only) or +11% for longer request formulation (Braschler and Ripplinger 2004). Similar mean performance differences have been found by Hedlund et al. (2004).

### 2.3 IR Models

In designing, implementing, and testing IR tools for European languages, different IR models have been used, such as variants of the vector-space models (Buckley et al. 1995; Manning et al. 2008), Okapi (Robertson et al. 2000), language models (Hiemstra 2000), and probabilistic approaches derived from "Deviation From Randomness" (DFR) (Amati and van Rijsbergen 2002). The formulations underlying these approaches are based on three main components, namely (1) the term frequency (*tf*) of the corresponding term in the document or the user's request, (2) the inverse document frequency (or *idf*), and (3) a length normalization procedure.

Essentially, these three factors encode the notion that a term should contribute most to the calculation of the item's score (or RSV, retrieval status value), if that term is found frequently in a document ("locally" frequent) and rarely in the overall collections ("globally" rare). The three factors have proven to be useful to discriminate between the major and minor semantic aspects of a document (or a request). Moreover, this formulation does not depend on the underlying

natural language which can be an Indo-European one (Savoy 2003a), an Indian language (Dolamic and Savoy 2010a), or even Chinese, Japanese, Korean (Savoy 2005), the last three requiring however a more complex tokenization procedure.

Overall, our experiments indicate that both Okapi and variants of DFR tend to produce the highest retrieval effectiveness over numerous languages using the CLEF test collections (composed mainly of newspapers), and are thus most “robust” towards the different characteristics of the languages we have studied. The IR schemes derived from a language model tend to produce high mean average precision, marginally lower than those achieved by the Okapi or some DFR approaches. In all these implementations however, the best values for the underlying parameters are not known in advance and may have an impact on the overall effectiveness.

### 3 Bilingual Information Retrieval

Bilingual Information Retrieval (BIR) corresponds to the simplest form of information retrieval in which the requests are written in one language and the information items in another. Often, the term “cross-language” (or “cross-lingual”) information retrieval (CLIR) is used as an alternative. The latter term is, however, less precise and can also be applied to scenarios with more than two languages involved. In nearly all cases, a direct matching between the query and the document surrogates does not work effectively in a bilingual scenario, and a translation stage must thus be incorporated during the IR process.

To achieve this, the simplest strategy is to translate the requests into the target language, knowing that queries are usually shorter than documents (query translation). The second approach consists of translating the whole text collection into the query language(s) (document translation). In this case, the translation process can be done off-line, and thus the translation process does not increase the response delay.

In some particular circumstances, the translation step can be ignored. Belonging to the same language family, some words may appear in different languages with the same or similar spelling (e.g., cognates such as, for example music, “Musik” (German), “musica” (Italian), “musique” (French), “música” (Spanish)). For some closely related languages, a rather large part of the vocabulary has similar spellings in the two languages, as for example, English and French, or German and Dutch. This aspect can be also explained by the presence of numerous loanwords (e.g., joy and “joie” (French)). Therefore, retrieval is possible when assuming that “English is simply misspelled French” (Buckley et al. 1997).

In this perspective for retrieval purposes, the translation stage is then replaced by a soft matching based on a spell corrector. This ingenious strategy is only possible for a limited number of closely related languages.

Moreover, this approach does not usually perform as well as an IR system with an explicit translation procedure (the solution achieves approximately 60% of the

effectiveness of a monolingual retrieval). In addition, sometimes the meaning differs even if the spelling looks similar (e.g., “demandes du Québec” must be translated into “requests of Quebec” and not as “demands of Quebec”).

Therefore, to achieve a good overall IR performance, a form of explicit translation must be included during the IR process. This can be achieved using various techniques as shown in Sect. 3.1. The next section presents an architecture based on a query-translation approach and indicates some effectiveness measures.

### 3.1 Translation Strategies

A good translation requires knowing the meaning of the source text, and therefore could be hard to perform perfectly automatically. Note, however, that in a retrieval scenario, it may not be necessary to render a translation in the classical sense. The role of the “translated” query is merely the retrieval of relevant items in the other language; for this, any representation of the query *intent* in the target language, whether directly recognizable as translation or not, is suitable.

During the translation process, different forms of ambiguity must be resolved. For example, the correct translation of a word or expression depends on the context (word sense disambiguation) as, for example, the translation of the word “bank” differs if one considers a river or a financial context. Similarly, the French word “temps” could be translated into “time,” “weather,” or even “tense”. Thus, for a given term, the translation process could be hard in one direction, but not in the other.

Moreover, not every word in one language does necessarily have a direct corresponding one in the target language (e.g., the occurrence of “have” in “have to” or “have” must usually be translated differently). Therefore, a word-by-word translation does not provide the best solution.

Multi-word expressions raise another set of ambiguities. Idiomatic expressions (e.g., “to see stars”) cannot be translated as is into the target language. In other cases, the culture generates expressions that do not have a direct equivalent in the target language (e.g., “a lame duck Congressman”).

As translation strategies (Zhou et al. 2012), the BIR experiments performed during the CLEF campaigns have tried different tools and IR models. As a first solution, one can use machine-readable bilingual dictionaries (MRD). In this case, each surface word in one language is searched in an MRD and the set of possible translations is returned. Even if some MRDs return, on average, only one or two possible translations, for some words the number of translations can be far larger (we have observed up to 15). It is not clear whether the IR system must take account only of the first one, the first  $k$  (with  $k = 3$  to 5), or simply all translations. Usually, the IR system assumes that the translations are provided in a rank reflecting their decreasing frequency or usefulness. Thus, a weight assigned to each translation can depend on its position in the returned list.

The issue of how many candidate translations for a term should be included in the translated query representation has been handled in different ways. Assuming that the MRD returns the candidate in descending order of frequency of occurrence, the output can be pruned by accepting at most  $k$  translation candidates. This approach is problematic if  $k > 1$ , since unambiguous source language terms will then be under-represented in the translated rendering. Hedlund et al. (2004) present a remedy to this with their “structuring of queries” approach, where the  $k$  translation candidates are weighted as a “synonym set”, instead of individually. They give results from experiments with three source languages (Swedish, Finnish, and German) and find consistent benefits of using structured queries. Greatest benefits are reported for Finnish, where they obtained an increase in retrieval effectiveness of up to 40%.

A more linguistically motivated alternative is the attempt to select “the” optimal translation candidate, e.g., through word sense disambiguation. Approaches using automatically generated dictionaries from corpora can be helpful here, as they can reflect specific domains in the context of which translation is less ambiguous. We will discuss relevant approaches to produce such statistical translation resources below.

MRDs as a translation tool must be integrated with caution. An MRD is not the same as a paper-based bilingual dictionary. In this latter case, each dictionary entry corresponds to a lemma (e.g., “to see”), but the surface word may include inflectional suffixes (e.g., “sees”, “saw”, “seen”). Thus, the link between the surface word and the lemma could be problematic.

Moreover, names may raise additional difficulties when they do not appear in the dictionary and sometimes the spelling varies from one language to the other (e.g., Putin, Poutine, Poetin). Of course, names are not limited to well-known politicians but can denote a product or an artwork (e.g., “Mona Lisa” (Italian), “La Joconde” (French) or “La Gioconda” (Spanish)). When names are relatively frequent, their translations can be obtained by consulting specialized thesauri (e.g., *JRC-Names*, *Arts and Architectures Thesaurus*, *The Getty Thesaurus of Geographic Names*). Similar data structures can also be built from other sources such as the *CIA World Factbook*, various gazetteers, or by downloading Wikipedia/DBpedia pages written in different languages. A similar solution can be applied to translate acronyms (e.g., UN must appear as ONU (in Spanish, French, Italian), UNO (in German), ONZ (in Polish), or YK (in Finnish)), under the assumption that a short sequence of uppercase letters corresponds to an acronym.

When a translation is not returned for a given word (out-of-vocabulary problem) resulting from a dictionary’s limited coverage, the usual reason is the presence of a name (e.g., London, Renault) and the corresponding word can be kept as it is or translated by the previously mentioned tools. In other cases, a word (corresponding to a name) should not be translated (e.g., Bush).

Finally, the most appropriate translation can depend on the national origin of the target collection. Each language is strongly related to a culture. Therefore, one word or expression can appear in a given region, not in another one (or with a different meaning). For example, the translation of “mobile phone” into French



can be “téléphone portable” (France), “téléphone mobile” (Belgium), “cellulaire” (Canada) or “natel” (Switzerland).

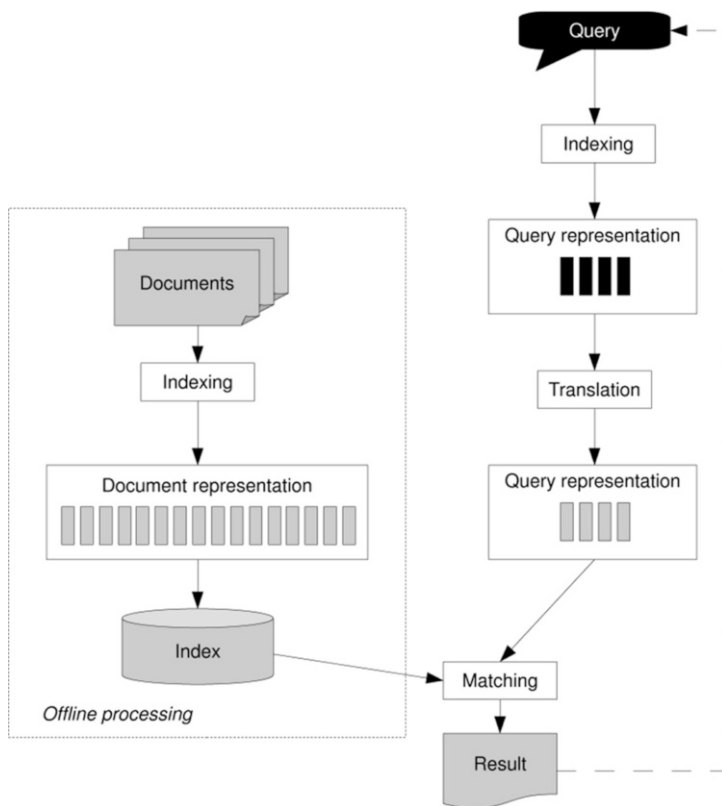
As a second translation strategy, one can adopt a machine translation (MT) system that will automatically provide a complete translation of a given request (or document) into the target language. As well-known examples, one can mention Google or Yahoo! online translation services. Various other systems have been developed such as Systran, Promt/Reverso, Babel Fish or WorldLingo. A classic example of the use of such automatic translation system is the Canadian weather forecast (started in 1971), while the latest version translates also weather warnings (Gotti et al. 2013).

As a third possibility of identifying proper translation candidates, we can apply a statistical translation model (Kraaij et al. 2003).<sup>2</sup> Advances in the effectiveness of machine translation systems reduce the role of statistical translation models for bilingual and multilingual retrieval to something of a niche role; however, there is still considerable potential for cases where special vocabulary (e.g., many proper names) and/or less frequently spoken languages are involved. Ideally, the model is built on the basis of a parallel corpus (i.e., a corpus that contains high-quality translations for all the documents) written in the desired languages. By aligning the translated documents at sentence level, pairs of terms across the languages are identified as translation candidates. Building a data structure from these pairs, the most probable match or the best  $k$  matches (Braschler and Schäuble 2001) can serve as retrieval terms.

In principle, this approach is workable independently of the languages considered. The availability of a suitable parallel corpus covering both the languages and the desired target domain, however, remains a concern. In Braschler (2004), we show that the requirement for a parallel corpus is not a strict one; instead, a comparable corpus that works on a much coarser “document similarity” basis, may be sufficient and may be much easier to obtain. Nie et al. (1999) discusses how suitable candidate documents can be identified in publicly accessible Web resources. Starting from a comparable corpus (Braschler 2004), shows how documents are “aligned” if they describe the same news event, even if produced independently by different authors. By modifying the  $tf\ idf$ -weighting formula to retrieve terms that co-occur in a training set of documents, a very large translation resource can be built that covers a vocabulary that is potentially much larger than that of MRDs (“similarity thesaurus”). Of course, the overall performance of such statistical translation systems depends on important factors, such as quality and size of the sources (Kraaij et al. 2003), along with the role played by cultural, thematic and time differences between the training corpora and the target domain.

---

<sup>2</sup>For example, by using the freely available Moses system, see [www.statmt.org/moses/](http://www.statmt.org/moses/).



**Fig. 1** Main architecture for a bilingual information retrieval system

### 3.2 Query Translation

To implement a query translation process, one can insert the automatic translation phase between the request acquisition and query indexing stage. As the request is usually rather short, the translation delay can be brief and done in real time at query time. As a translation strategy, one can implement an approach based on MRDs, an MT system or using a statistical translation model. Based on our experiments, the MT approach tends to produce the highest average performance level. As a variant depicted in Fig. 1,<sup>3</sup> the query representation can be generated in the query language and then translated into the target language in which the search is performed. If needed, the search result can be translated into the query language.

The overall quality of a translation system depends also on the language pair, and having English as one of the two languages tends to produce better results

<sup>3</sup>The figures appearing in this chapter are reproduced from Peters et al.'s book (Peters et al. 2012).

(the demand for automatic translation from/to English is clearly higher than for another language). In comparing different languages when using the Google system (Dolamic and Savoy 2010b), we observe that the translation from queries written in French or Spanish in order to search in an English collection was easier than it was from the Chinese. Based on a DFR model, the MAP obtained for the bilingual search using the French or Spanish language in the query language achieves 92% of the MAP obtained for the monolingual search. This value decreases to 90% with German topics, and 82% with simplified Chinese as language. With the Yahoo! translation service, the situation was somewhat comparable, with the French language achieving the best MAP (82% of the monolingual search), and using Chinese as the query language was the most difficult (only 56% for the monolingual search).

As the first source of translation errors, one can find the problem of polysemy and synonymy attached to a word. With the French request “Vol du cri” (“Theft of *The Scream*”), the word “vol” can be translated into “flight” or “theft”, both with a high probability of being correct. In other cases, the choice in the target language seems irrelevant from a semantic point of view because two words are viewed as synonyms (e.g., the German word “Wagen” could be translated into “car” or “automobile”). From an IR perspective, one of these possible correct translations will provide more relevant items (e.g., car) than the other (e.g., automobile).

The second main source of translation errors comes from names. For example, in the request “Death of Kim Il Sung”, the last word can be incorrectly analyzed as the past participle of the verb “to sing”. Therefore, the returned translation is inappropriate to retrieve all pertinent information items. With another translation tool, the term “Il” was incorrectly recognized as the chemical acronym for Illinium (an discontinued chemical element). Finally, the Spanish word “El Niño” must not be translated into English (i.e. “the boy”) but must be kept as is when the underlying domain concerns global warming. Of course, manual translation does not guarantee correct expressions.<sup>4</sup>

In order to limit translation ambiguity, one can automatically add terms to the submitted request before translating it into the target language (Ballesteros and Croft 1997). In this case, the query is first used to search within a comparable collection of documents written in the request language. Based on a pseudo-relevance feedback scheme, new and related terms can then be added to the query before translation. Such new terms may reflect morphological variations (e.g., from a query about “London”, the extended query may include additional terms or related concepts such as “Britain, British, PM, England”).

As a second strategy to improve the BIR system, the translation stage can take account of more than one translation approach or source. It was shown that combining multiple translation sources (Savoy 2004) tends to improve the overall retrieval effectiveness (Savoy and Berger 2005). For example, using queries written

---

<sup>4</sup>In a hotel cloakroom in Germany, the following faulty translation was found: “Please hang yourself here.” (Crocker 2006).

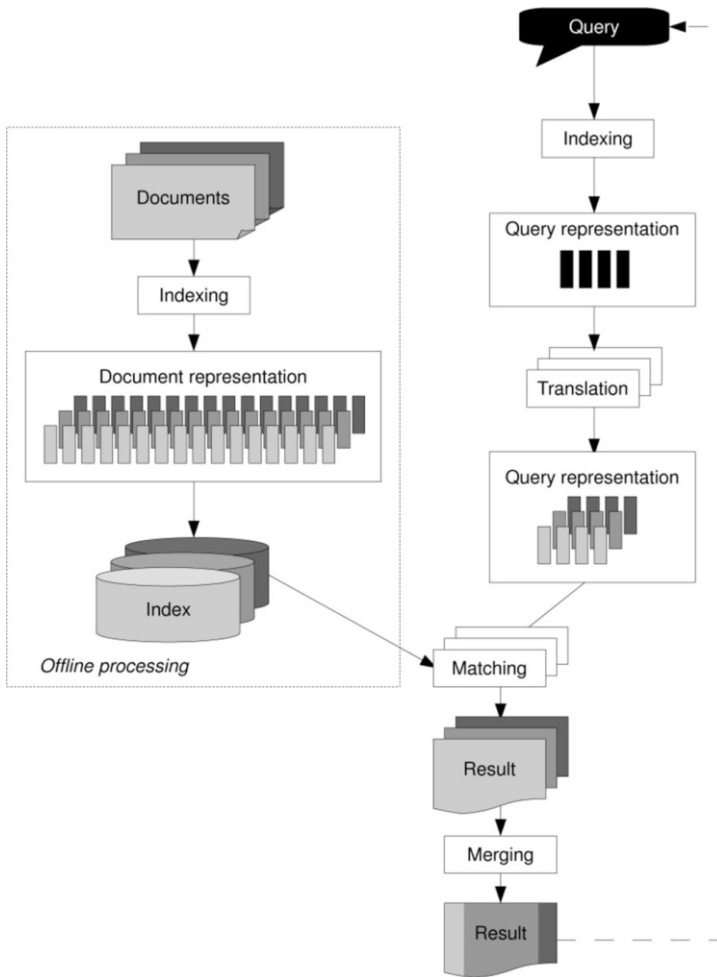
in English to search a collection written in another language, we have combined two alternative translated representations of the query. In the best case, searching in a French collection, the MAP can be improved from 8% to 12% compared to a single translation. Similar average enhancements can be found using the Spanish or Russian language (Savoy 2004). With the Italian language, the improvement was even higher, from 18% to 30%. When compared to the corresponding monolingual search and combining two translation tools, the performance difference is similar when searching in the French corpus (with English requests), with a 8% decrease for a collection written in German and around 10% decrease for the Spanish or Italian language. Those performance levels can be achieved when having the English as one of the languages. Of course, such a translation strategy is clearly more complex to design and to maintain in a commercial environment.

## 4 Multilingual Information Retrieval

Designing effective Multilingual Information Retrieval (MIR) systems corresponds to a very challenging issue. In such a context, the request can be written in one language while the information items appear in many languages. As for BIR, the translation process must be included in the IR process generating an additional level of uncertainty. In such an IR system, we usually assume that one document collection corresponds to one language. Therefore, the search must be done across different separate collections or languages. However, an MIR system can be built with different architectures, and the simplest one is based on a query-translation approach as described in Sect. 4.1. More complex approaches, usually achieving better retrieval effectiveness, implement a document translation phase as discussed in Sect. 4.2 or both a document and query translation process as described in Sect. 4.3.

### 4.1 *Multilingual IR by Query Translation*

As a first MIR architecture, one can simply translate the submitted request into all target languages. Note, however, that this approach suffers from scaling issues: as the number of languages to be covered grows, so does the number of translated representations that need to be produced. The number of bilingual language pairs can thus quickly become prohibitively large. After producing the individual translations, the search is performed separately in each language (or collection), each returning a ranked list of retrieved items. MIR then presents an additional problem. How can one merge these results to form a single list for the user in an order reflecting the pertinence of the retrieved items, whatever the language used (“merging”)? Figure 2 depicts the overall MLIR process based on a query translation (QT) strategy.



**Fig. 2** Main architecture for a query translation model for a cross-language information retrieval system

As a first merging approach, one might assume that each language contains approximately the same number of pertinent items and that the distribution of relevant documents is similar across the result lists. Using the rank as the sole criteria, the simplest solution is then to interleave the retrieved records in a round-robin fashion. As an alternative, one can suggest a biased round-robin approach which extracts not one document per collection per round but one document for each smaller collections and more than one for larger ones (Braschler et al. 2003).

To account for the document score (or RSV) computed for each retrieved item (or the similarity value between the retrieved record and the query), one can formulate the hypothesis that each collection is searched by the same or a very similar

search engine. In such cases, the similarity values are directly comparable across languages/collections. Such a strategy, called raw-score merging, produces a final list sorted by the document score computed separately by each collection.

However, as demonstrated by Dumais (1994), collection-dependent statistics in document or query weights may vary widely among collections, and therefore this phenomenon may invalidate the raw-score merging hypothesis. But different evaluations carried out using English only documents have demonstrated that the raw-score merging strategy sometimes leads to satisfactory performance (Rasolofoa et al. 2003).

As a third merging strategy, one can normalize document scores within each collection by dividing them by the maximum score (i.e. the document score appearing in the first position (Fox and Shaw 1994), a strategy denoted “Norm Max”). This procedure could generate more comparable document scores across all languages/collections. As a variant of this normalized score merging scheme, Powell et al. (2000) suggest normalizing the document scores by taking the maximum and minimum document score (approach denoted “Norm RSV”) and explained by Eq. (1).

$$Norm\ RSV(D_k^i) = \frac{RSV_k^i - Min(RSV^i)}{Max(RSV^i) - Min(RSV^i)} \quad (1)$$

where  $RSV_k^i$  indicates the retrieval score of document  $k$  in the  $i$ th retrieved list, and  $Max(RSV^i)$  (respectively  $Min(RSV^i)$ ) the maximum (minimum) RSV value appearing in the  $i$ th list.

As a fifth merging strategy, the “Zscore” approach (Savoy 2004) has been suggested in which the normalization of the RSV values depends on the RSV distribution, using its mean ( $Mean(RSV^i)$ ) and estimated standard deviation ( $Std(RSV^i)$ ). The precise definition is provided by Eq. (2).

$$Z\ score(D_k^i) = \frac{RSV_k^i - Mean(RSV^i)}{Std(RSV^i)} + \delta^i \quad \delta^i = \frac{Mean(RSV^i) - Min(RSV^i)}{Std(RSV^i)} \quad (2)$$

Finally, machine learning methods can be applied to improve the merging operation. In this perspective, a logistic regression approach can be used to estimate the probability of relevance for a given document, based on its retrieval status value and the natural logarithm of its rank. The final list is sorted according to these estimates. The evaluation is performed based on the leaving-one-out evaluation strategy producing an unbiased estimator of the real performance.

To analyze the quality of these merging operators, the CLEF 2004 test collection has been selected (Savoy and Berger 2005). This corpus contains newspapers articles written in English, French, Finnish, and Russian. Table 1 indicates the number of queries with relevant items in each language, as well as the MAP achieved when applied to the original queries (column denoted “Manual” or monolingual run).

**Table 1** MAP of each single run

| Query (TD) |                   | Mean average precision (MAP) |             |             |
|------------|-------------------|------------------------------|-------------|-------------|
| Language   | Number of queries | Manual                       | Condition A | Condition B |
| English    | 42                | 0.5580                       | 0.5580      | 0.5633      |
| French     | 49                | 0.4685                       | 0.4098      | 0.4055      |
| Finnish    | 45                | 0.4773                       | 0.2956      | 0.2909      |
| Russian    | 34                | 0.3800                       | 0.2914      | 0.2914      |

**Table 2** MAP of various multilingual merging strategies

| Query (TD)          | Mean average precision (MAP) |             |            |
|---------------------|------------------------------|-------------|------------|
| Merging operator    | Condition A                  | Condition B | Difference |
| Round-robin         | 0.2386                       | 0.2358      | -1.2%      |
| Biased round-robin  | 0.2639                       | 0.2613      | -1.0%      |
| Raw-score           | 0.0642                       | 0.3067      | 377.7%     |
| Norm max            | 0.2552                       | 0.2484      | -2.7%      |
| Norm RSV            | 0.2899                       | 0.2646      | -8.7%      |
| Z-score             | 0.2669                       | 0.2867      | 7.4%       |
| Logistic regression | 0.3090                       | 0.3393      | 9.8%       |
| Optimal selection   | 0.3234                       | 0.3558      |            |

Under Condition A (bilingual runs with English queries), we have tried to obtain a high MAP per language, applying different IR models with distinctive parameter values for each language. Under Condition B, the same IR model (a variant of the DFR family) is used for each language (with similar parameter values). This last choice reflects the case where a single IR model is used to search across different collections/languages.

Table 2 reports the MAP achieved when applying different merging operators. The round-robin method must be viewed more as a baseline than a really effective approach. When distinct IR models are merged (Condition A), the raw-score merging strategy resulted in poor retrieval effectiveness. On the other hand, when applying the same IR model (with similar parameter values), the raw-score approach offers higher MAP. The normalization procedures (either by the Norm Max or the Norm RSV) or the Z score technique tend to produce better retrieval results than the round-robin technique under both conditions.

In some circumstances, an effective ranking can be learnt from past results. As an example, a logistic regression model can use both the rank and the document score as explanatory variables to predict the probability of document relevance. When such training sets are available and the similarity between trained and test topics is high, the merging achieved can be significantly better than the round-robin merging as well as better than the simple normalization approaches (see Table 2). Finally, the last row of Table 2 reports the optimal merging result that can be achieved based on the returned lists per language. Compared to the round-robin strategy, this optimal merging offers a 36% improvement under Condition A (0.3234 vs. 0.2386) and +50% under Condition B (0.3558 vs. 0.2358).

## 4.2 Document Translation

Document translation (DT) provides an attractive alternative approach avoiding the merging problem. By translating all documents into a single, unified target language, the multilingual retrieval problem is essentially reduced to a monolingual one. Interestingly, the merging problem is thus avoided altogether. For reasons of its superior language resources, a pertinent choice for the pivot language is English. To justify this choice, we describe the following experiment.

In the experiment (Savoy and Dolamic 2010), we needed to translate 299 queries written in German to search in a French collection. Compared to a monolingual run (MAP: 0.6631), the achieved MAP was 0.4631 resulting in a decrease of around 30%. Using English as the query language, the MAP was 0.5817, for a performance difference of 12% compared to the monolingual run. Clearly the translation quality was higher from English than from the German language. Moreover, we need to limit the number of translation pairs. In our case, we are using English as pivot language. In a second stage, we first translate the German queries into English and then into French. After this two-stage translation, it is reasonable to expect a poor retrieval performance. Using English as pivot language, the resulting MAP was 0.5273, with an average decrease of only 20% (compared to the 30% with a direct translation from German to French). Similar good retrieval performances with a pivot language were observed in Hedlund et al. (2004). An example of the resulting MLIR process is depicted in Fig. 3.

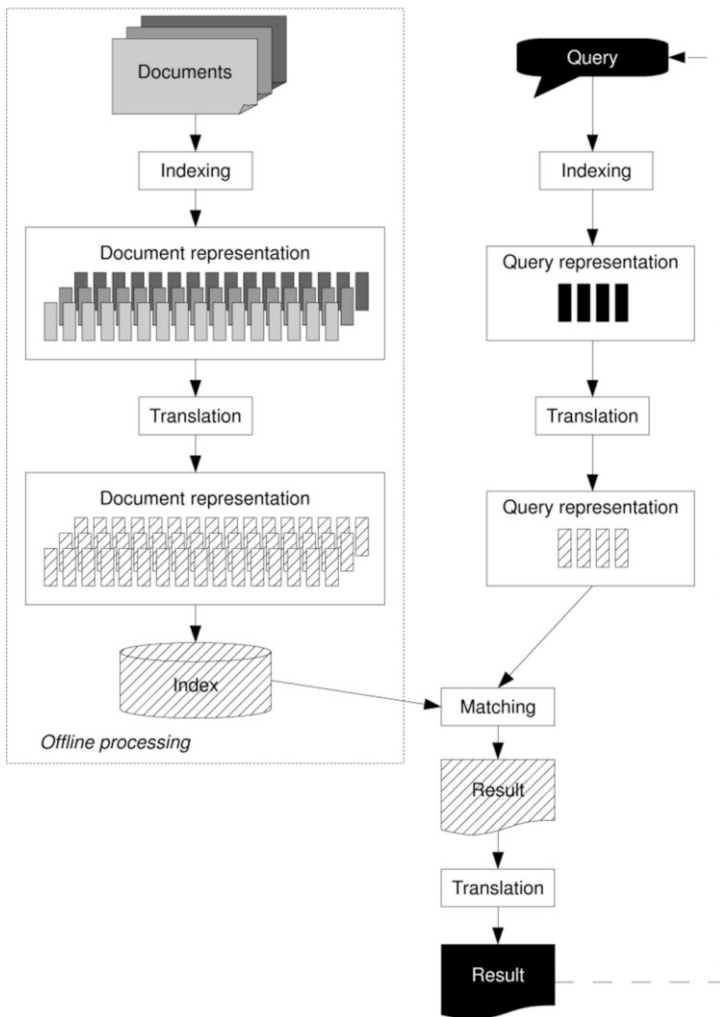
As a second model, we can translate all text collections into all query languages. Receiving the query in one of the available languages, the search is then performed as a monolingual one. In this case, no translation is performed during query processing.

All translation strategies outlined in Sect. 3.1 equally apply to document translation. Since a document (retrievable item) is typically much longer than a query, more context is available, and problems with out-of-vocabulary terms and synonymy tend to be less pressing. Moreover, there is justified hope that the information contained in some of the untranslatable terms is represented, at least partially, in the remainder of the document. Note that, analogously to the situation in query translation, a translation in the “classical sense” is not necessary; any rendering of the document into a representation in the target language that is suitable for retrieval will do (e.g., the syntax of the target language is not always perfectly respected (“pseudo translation”)).

Translation of large document collections, even if automated, is a costly task. The document translation approach also does not scale well as the number of query languages grows—in essence, the collection has to be replicated (and re-translated) for each target language. On the positive side, it is possible to do this translation offline, with no performance impact of translation during query time.

Examples of document translation-based experiments in the CLEF ad hoc tracks are reported in Braschler (2004) and McNamee and Mayfield (2002, 2004). In our experiments, we have gained the most insight in document translation behavior





**Fig. 3** Main architecture for a document translation model for a cross-language information retrieval system

from using the CLEF 2002 test collection containing documents written in English, German, French, Italian, and Spanish. We have used the German query set.

In order to have an idea about the performance differences between a QT and DT approaches, we have conducted the following experiment. First, we considered two query translation (QT) approaches, namely round-robin, and biased round-robin. As shown in Table 3, these two QT approaches tend to produce similar overall mean average precision. In the last column, we have indicated the performance difference with the round-robin solution.

**Table 3** Eurospider experiments on the CLEF 2002 multilingual corpus, German queries

| Strategy                              | Mean average precision | Difference |
|---------------------------------------|------------------------|------------|
| Query translation, round-robin        | 0.3249                 |            |
| Query translation, biased round-robin | 0.3369                 | +3.7%      |
| Document translation                  | 0.3539                 | +8.9%      |
| Optimal selection                     | 0.4876                 | +50.1%     |

**Table 4** Hybrid approach vs. document translation only or query translation only

| Strategy                  | Mean average precision (MAP) |           |           |
|---------------------------|------------------------------|-----------|-----------|
|                           | CLEF 2000                    | CLEF 2001 | CLEF 2002 |
| Query translation (QT)    | 0.2500                       | 0.2773    | 0.2876    |
| Document translation (DT) | 0.2816                       | 0.3099    | 0.3539    |
| DT + QT                   | 0.3107                       | 0.3416    | 0.3554    |

Second, in the fourth row, our document translation (DT) is evaluated. One can see that this DT approach outperformed the three QT strategies. However, when comparing to the benchmark of “Optimal Selection” (see Sect. 4.1), i.e. under the condition that the merging problem is “solved”, a different conclusion must be drawn. Note, however, that compared to simple merging strategies, we have found consistently better results for document translation across all years where we have made such comparisons (CLEF 2000–2002) as reported in Table 4.

### 4.3 Hybrid Approaches

Using the mean as a measure, we obtain a synthetic value reflecting the overall performance of an IR system. The differences between the average precision achieved by each query are however hidden. Looking at individual queries, it becomes evident that performance differences between query translation and document translation approaches vary greatly. To take advantage of both translation models, a hybrid approach can combine their outputs. In this scenario, a more robust solution can be proposed with respect to outliers. Indeed, our experiments on CLEF 2000–2002 test collections have shown an increase in mean average precision for all three years as reported in Table 4. As indicated previously, the document translation strategy performs better than the query translation approach over the three years. When comparing the document translation (second row) with the hybrid model (last row), the performance differences are always in favor of the hybrid model, although the difference for 2002 is negligible.

Analyzing query-by-query these results, we can see that the hybrid strategy proposes a better average precision for the majority of the queries. In Table 5, we have depicted the number of queries performing better in terms of average precision (over the set of 50 queries available each year). For example, for the CLEF 2001 collection, 41 out of 50 queries benefit from the hybrid approach when compared to

**Table 5** Impact on individual queries

| Strategy            | CLEF 2000 | CLEF 2001 | CLEF 2002 |
|---------------------|-----------|-----------|-----------|
| DT + QT vs. DT only | 32:8      | 41:9      | 28:22     |
| DT + QT vs. QT only | 31:9      | 36:14     | 41:9      |

document translation only, while this value reaches 36 when comparing to the query translation approach.

## 5 Conclusion

During our ten years of participation in the mono-, bi-, and multilingual tracks at CLEF, we have designed, implemented, and evaluated various IR tools for a dozen of European natural languages. Those experiments tend to indicate that the IR models validated on various English collections (e.g., TREC, NTCIR, CLEF, INEX) perform also very well with other European (Savoy 2003a), Indian (Dolamic and Savoy 2010a), or Far-East (Savoy 2005) languages. No special adaptation is really required when considering the *tf*, *idf*, and length normalization components. On the other hand, some IR procedures must take into account the specifics of each language.

Each natural language presents its own difficulties when building effective IR systems. To generate a stopword list, we suggest considering all closed part-of-speech categories (determiners, prepositions, conjunctions, pronouns, and auxiliary verb forms). In this list, an inspection is needed to verify, according to the target application or domain, whether some forms must be removed or not from the stopword list (e.g., the article “a” can appear in the context of “vitamin A”).

To develop a stemmer for a new language, we suggest focusing mainly on morphological variations related to nouns and adjectives, and to ignore the usually too numerous suffixes related to verbs. Moreover, removing only the inflectional suffixes seems to be good practice for many languages. Adopting this approach, the edit distance between the search term introduced by the user and its internal representation is rather small. With a light stemmer, one can improve the MAP in the range of 5% to 10% (e.g., French or German language) up to 96% (Russian).

If needed, and according to the target application, an advanced stemmer can be proposed to remove both inflectional and derivational suffixes. The enhancement over a light stemmer is between  $-1\%$  (Russian) to  $+6\%$  (French). Trying to remove verbal suffixes tends to be more problematic by generating too many incorrect confluences for nouns and adjectives. For the German language only, we recommend implementing an automatic decomposing procedure, leaving both the compound and its separate components in the document or query surrogate. This strategy can increase the mean performance by 23% (Braschler and Ripplinger 2004).

Recent research has been conducted to analyze in a more systematic way the effect of different stopword lists and stemmers, as well as their combined effect (Ferro and Silvello 2016a,b).

When implementing a bilingual IR system, the crucial component is clearly the translation procedure. When the pair of languages includes English and one of the most widely spoken languages (such as Spanish, German, or French), currently available machine-translation systems offer high effectiveness from an IR point of view (Dolamic and Savoy 2010b). Even if the translation is not fully correct from a linguistic standpoint, the search engine is able, on average, to find the appropriate related search terms and to retrieve the pertinent items. In such circumstances, the decrease of the mean performance compared to the monolingual setting is rather limited ( $-5\%$  to  $-12\%$ ), and in the best case, no degradation occurs. For other languages (e.g., Finnish, Polish), the number of translation tools is rather limited and their quality is clearly inferior to those available for the most frequently spoken languages. The retrieval performance can however be improved by combining multiple translations of the same texts on the one hand, and on the other, by applying some query expansion before the translation. However, such IR strategies render the final system more complex and difficult to maintain.

When the translation resources available are limited or absent, the usual solution is to generate a statistical translation system based on parallel corpora (Kraaij et al. 2003). In this case, the mean retrieval precision typically decreases substantially (from 10% to 40%). Finally, more specific IR models have been proposed to take account of the additional uncertainty generated by the translation process.

Multilingual IR corresponds to our most complex situation in which the overall performance depends on many factors and where the quality of the translation plays an important role. From an architecture point of view, two main approaches have been tested. The simplest one is based on query translation (QT) in which the submitted query is translated automatically into all the target languages. The search is then done separately in all languages, and the results are then merged to generate a single ranked list of retrieved items to be presented to the user. The main difficulty in this model is the merging process that can substantially degrade the overall performance. Our experiments indicate that selecting a form of normalization of the document score (e.g., Norm RSV or the Z score) can offer a reasonable overall IR performance.

In a document translated (DT) model, all documents are translated into a single pivot language (usually in English). The submitted request is also automatically translated into this pivot language. The search process is then done in a single language and the resulting ranked list can be directly returned to the user. Such solutions tend to produce a better overall retrieval performance compared to query translation approaches.

**Acknowledgement** The authors would like to thank the CLEF organizers for their efforts in developing the CLEF test collections.

## References

- Amati G, van Rijsbergen CJ (2002) Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans Inf Syst* 20:357–389
- Ballesteros L, Croft BW (1997) Phrasal translation and query expansion techniques for cross-language information retrieval. In: *Proceedings ACM SIGIR*. ACM Press, New York, pp 84–91
- Braschler M (2004) Combination approaches for multilingual text retrieval. *Inform Retrieval J* 7:183–204
- Braschler M, Ripplinger B (2004) How effective is stemming and compounding for German text retrieval? *Inform Retrieval J* 7:291–316
- Braschler M, Schäuble P (2001) Experiments with the europsider retrieval system for CLEF 2000. In: Peters C (ed) *Cross-language information retrieval and evaluation*. LNCS, vol 2069, Springer, Berlin pp 140–148
- Braschler M, Göhring A, Schäuble P (2003) Europsider at CLEF 2002. In: Peters P, Braschler M, Gonzalo J, Kluck M (eds) *Advances in cross-language information retrieval: third workshop of the cross-language evaluation forum (CLEF 2002) revised papers*. LNCS, vol 2785. Springer, Berlin, pp 164–174
- Buckley C, Singhal A, Mitra M, Salton G (1995) New retrieval approaches using SMART. In: *Proceedings TREC-4*, NIST, Gaithersburg, pp 25–48
- Buckley C, Singhal A, Mitra M, Salton G (1997) Using clustering and superconcepts within SMART: TREC-6. In: *Proceedings TREC-6*, NIST, Gaithersburg, pp 107–124
- Chen A (2004) Report on CLEF-2003 monolingual tracks: fusion of probabilistic models for effective monolingual retrieval. In: Peters C, Gonzalo J, Braschler M, Kluck M (eds) *Comparative evaluation of multilingual information access systems*, LNCS, vol 3237. Springer, Berlin, pp 322–336
- Crocker C (2006) *Løst in Tränslatioπ. Misadventures in English abroad*. Michael O’Mara Books, London
- Dolamic L, Savoy J (2009a) Indexing and searching strategies for the Russian language. *J Am Soc Inf Sci Technol* 60:2540–2547
- Dolamic L, Savoy J (2009b) Indexing and stemming approaches for the Czech language. *Inf Process Manag* 45:714–720
- Dolamic L, Savoy J (2010a) Comparative study of indexing and search strategies for the Hindi, Marathi and Bengali languages. *ACM Trans Asian Lang Inf Process* 9(3):11
- Dolamic L, Savoy J (2010b) Retrieval effectiveness of machine translated queries. *J Am Soc Inf Sci Technol* 61:2266–2273
- Dolamic L, Savoy J (2010c) When stopword lists make the difference. *J Am Soc Inf Sci Technol* 61:200–203
- Dumais ST (1994) Latent semantic indexing (LSI) and TREC-2. In: *Proceedings TREC-2*, vol #500-215. NIST, Gaithersburg, pp 105–115
- Fautsch C, Savoy J (2009) Algorithmic stemmers or morphological analysis: an evaluation. *J Am Soc Inf Sci Technol* 60:1616–1624
- Ferro N, Silvello G (2016a) A general linear mixed models approach to study system component effects. In: *Proceedings ACM SIGIR*. ACM Press, New York, pp 25–34
- Ferro N, Silvello G (2016b) The CLEF monolingual grid of points. In: Fuhr N, Quaresma P, Gonçalves T, Larsen B, Balog K, Macdonald C, Cappellato L, Ferro N (eds) *Experimental IR meets multilinguality, multimodality, and interaction. Proceedings of the eighth international conference of the CLEF association (CLEF 2017)*. LNCS, vol 9822. Springer, Berlin, pp 13–24
- Fox C (1990) A stop list for general text. *ACM-SIGIR Forum* 24:19–35
- Fox EA, Shaw JA (1994) Combination of multiple searches. In: *Proceedings TREC-2*, vol 500-215. NIST, Gaithersburg, pp 243–249
- Gotti F, Langlais P, Lapalme G (2013) Designing a machine translation system for the Canadian weather warnings: a case study. *Nat Lang Eng* 20:399–433

- Harman DK (1991) How effective is suffixing? *J Am Soc Inf Sci* 42:7–15
- Hedlund T, Airio E, Keskkustalo H, Lehtokangas R, Pirkola A, Järvelin K (2004) Dictionary-based cross-language information retrieval: learning experiences from CLEF 2000–2002. *Inf Retrieval J* 7:99–120
- Hiemstra D (2000) Using language models for IR. PhD thesis, CTIT, Enschede
- Kraaij W, Nie JY, Simard M (2003) Embedding web-based statistical translation models in cross-lingual information retrieval. *Comput Linguist* 29:381–419
- Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge
- McNamee P, Mayfield J (2002) Scalable Multilingual Information Access. In: Peters P, Braschler M, Gonzalo J, Kluck M (eds) *Advances in cross-language information retrieval*. LNCS, vol 2785. Springer, Berlin, pp 207–218
- McNamee P, Mayfield J (2004) Character N-gram tokenization for European language text retrieval. *Inf Retrieval J* 7:73–98
- McNamee P, Nicholas C, Mayfield J (2009) Addressing morphological variation in alphabetic languages. In: *Proceedings ACM - SIGIR*. ACM Press, New York, pp 75–82
- Moulinier I (2004) Thomson legal and regulatory at NTCIR-4: monolingual and pivot-language retrieval experiments. In: *Proceedings NTCIR-4*, pp 158–165
- Nie JY, Simard M, Isabelle P, Durand R (1999) Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In: *Proceedings ACM - SIGIR*. ACM Press, New York, pp 74–81
- Paik JH, Parai SK (2011) A fast corpus-based stemmer. *ACM Trans Asian Lang Inf Process* 10(2):8
- Paik JH, Parai SK, Dipasree P, Robertson SE (2013) Effective and robust query-based stemming. *ACM Trans Inf Syst* 31(4):18
- Peters C, Braschler M, Clough P (2012) Multilingual information retrieval. From research to practice. Springer, Berlin
- Porter MF (1980) An algorithm for suffix stripping. *Program* 14:130–137
- Powell AL, French JC, Callan J, Connell M, Viles CL (2000) The impact of database selection on distributed searching. In: *Proceedings ACM-SIGIR*. ACM Press, New York, pp 232–239
- Rasolofy Y, Hawking D, Savoy J (2003) Result merging strategies for a current news metasearcher. *Inf Process Manage* 39:581–609
- Robertson SE, Walker S, Beaulieu M (2000) Experimentation as a way of life: Okapi at TREC. *Inf Process Manage* 36:95–108
- Sanders RH (2010) German, biography of a language. Oxford University Press, Oxford
- Savoy J (2003a) Cross-language information retrieval: experiments based on CLEF 2000 corpora. *Inf Process Manage* 39:75–115
- Savoy J (2003b) Cross-language retrieval experiments at CLEF 2002. In: Peters P, Braschler M, Gonzalo J, Kluck M (eds) *Advances in cross-language information retrieval*. LNCS, vol 2785. Springer, Berlin, pp 28–48
- Savoy J (2004) Combining multiple strategies for effective monolingual and cross-lingual retrieval. *Inf Retrieval J* 7:121–148
- Savoy J (2005) Comparative study of monolingual and multilingual search models for use with Asian languages. *ACM Trans Asian Lang Inf Process* 4:163–189
- Savoy J (2006) Light stemming approaches for the French, Portuguese, German and Hungarian languages. In: *Proceedings ACM-SAC*. ACM Press, New York, pp 1031–1035
- Savoy J (2008a) Searching strategies for the Bulgarian language. *Inf Retrieval J* 10:509–529
- Savoy J (2008b) Searching strategies for the Hungarian language. *Inf Process Manage* 44:310–324
- Savoy J, Berger PY (2005) Selecting and merging strategies for multilingual information retrieval. In: Peters C, Clough P, Gonzalo J, Jones GJF, Kluck M, Magnini B (eds) *Multilingual information access for text, speech and images*. LNCS, vol 3491. Springer, Berlin, pp 27–37
- Savoy J, Dolamic L (2010) How effective is Google’s translation service in search? *Commun ACM* 52:139–143
- Zhou D, Truran M, Brailsford T, Wade V, Ashman H (2012) Translation techniques in cross-language information retrieval. *ACM Comput Surv* 45(1):1

# The Challenges of Language Variation in Information Access



**Jussi Karlgren, Turid Hedlund, Kalervo Järvelin, Heikki Keskustalo, and Kimmo Kettunen**

**Abstract** This chapter will give an overview of how human languages differ from each other and how those differences are relevant to the development of human language understanding technology for the purposes of information access. It formulates what requirements information access technology poses (and might pose) to language technology. We also discuss a number of relevant approaches and current challenges to meet those requirements.

## 1 Linguistic Typology

Information access technology—such as information retrieval and related applications—is largely about finding and aggregating meaning from human language, and mostly, so far, from text. On a superficial level, it may seem as if human languages vary a great deal, but they are in fact similar to each other, especially in written form: they share more features than differences. What meaning is and by which means it is encoded in human text is a contentious research topic in itself, but that there is meaning in human utterances and that it is systematically recoverable is not.

The number of languages in the world is difficult to assess, but is usually put at being around 7000. More than 90% of those languages are spoken by populations of

---

J. Karlgren (✉)  
Gavagai and KTH Royal Institute of Technology, Stockholm, Sweden  
e-mail: [jussi@gavagai.se](mailto:jussi@gavagai.se)

T. Hedlund  
Helsinki, Finland

K. Järvelin · H. Keskustalo  
University of Tampere, Tampere, Finland

K. Kettunen  
The National Library of Finland, Helsinki, Finland

less than a million and more than half of them by language communities numbering less than 10,000. Many of those languages—primarily the smaller ones—are falling out of use, with some estimates putting about half of the world's languages at risk of disappearing. The number of speakers is unevenly distributed: at the other end of the scale the twelve or so largest languages cover half of the population of the world (Lewis et al. 2009; Dryer and Haspelmath 2011). The details of these facts of course depend crucially on how one language is demarcated from another, which is non-trivial, depending not only on linguistics but also on politics and geography. The variation between human languages is studied in the field of *linguistic typology*, which studies both systematic differences and likenesses between languages (Velupillai 2012).

Such variation between human languages is first, and most obviously, evident in their writing systems. Some languages use some variation of phonetic writing such as alphabetic or syllabic systems; other systems are based on ideograms; some separate tokens by whitespace, some do not. Some writing systems omit what others require: semitic languages usually do not include vowels, for instance. This type of variation is mostly superficial and is no longer a major challenge for information systems. More importantly, only about half of the world's languages are ever written at all and thus not accessible to most of today's information systems. However, the practical challenge of accommodating various writing systems, character sets, and their encodings, in view of many coexisting and legacy standards may still impact performance.

Secondly, human languages vary in the way they organise the referents the speaker communicates about into a coherent utterance. Some languages impose strict requirements on the order of the constituents of a clause, making use of *word order* an obligatory marker; others allow permutations of constituents within an utterance without much meaning change. Some languages render words in different forms through more or less elaborate *inflection* systems, depending on what role they play in the utterance; others let words appear in more or less invariant form. These two aspects of variation—inflection and word order—are in the most general sense in a trade-off relation: languages with strict word order tend to have less complex systems for inflection.

Thirdly, many languages combine words or bits of words to make larger words or *compounds* or *derivations*; others prefer to keep words or meaningful units separate.

Fourthly, information that is obligatory to include for some languages may be optional or not mentioned in others.

On another level of abstraction, *genres* and various cultural factors influence which topics are discussed and in which terms. The variation is even more evident with the advent of *new text types*.

We will return to all of the above variational dimensions in turn. More generally, however, all human languages share important features. Languages are *sequential*: they consist of sequences of meaning-bearing units which combine into useful utterances of salience to their speaker and author, and mostly of interest to their intended or unintended audience. Languages are *referential*: the utterances are composed of expressions which refer to entities, processes, states, events, and their



respective qualities in the world. Languages are *compositional*: the constituents of utterances combine to a meaningful whole through processes which to some extent are general and to some extent are bound to situation, context, and participants.

And in the end information access is all about meaning. In the case of text retrieval, about the semantics of a text and the utterances in it.

## 2 Requirements from Application Domains

The focus of information retrieval experiments has been on the use case of *ad hoc information retrieval*: the process whereby a concise expression of information need is exchanged for a set or ranked list of documents or other information items. To achieve levels of performance in every or most languages to match the level that systems achieve in English and other widely used languages with large speaker populations, more analysis of the target language is often necessary. This is even more true when the use case is extended to Cross-Language Information Retrieval (CLIR), where a query in one language is expected to deliver results in other languages, possibly in combination with results in the target language.

Other related tasks, ranging from media monitoring and routing to sentiment analysis to information extraction often require more sophisticated models and typically more processing and analysis of the information items of interest. Much of that processing needs to be aware of the specifics of the target languages.

Mostly, the various mechanisms of variation in human language pose *recall challenges* for information systems. Texts may treat a topic of interest but use linguistic expressions which do not match the expectations of the system or the expression of information need given by the user: most often due to vocabulary mismatch. This is especially true for users who may know the target language only to some extent, and who may not be able to specify their information need with as much finesse as native language users would: the benefits of query translation in web search benefits those with poor to moderate competence in the target language more than those who are fluent. Since CLIR will in such cases rely on translating an information need from a source language to a target language, the quality of the translation dictionary or service is a crucial factor for the quality of the end result, whether the translation is done at query time or at indexing time (Airio 2008).

Translation is not always possible between arbitrary language pairs, due to lack of resources: see e.g. Rehm and Uszkoreit (2012) for an overview of what resources are available. In such cases, a transitive approach can be adopted, where translation is done from language  $A$  to language  $B$  by way of translation via a *pivot language*  $C$ , if translation resources or services for  $A \iff B$  are unavailable but can be found for  $A \iff C$  and  $C \iff B$ . This obviously risks inducing a level of noise and spurious translation candidates, but has been shown to work adequately in many task scenarios (Gollins and Sanderson 2001; Lehtokangas et al. 2004).

## 2.1 *Cultural Differences and Differences in Genre Repertoire*

On the highest level of abstraction, differences between cultural areas are often reflected in how a topic is treated in linguistic data. This may not seem a challenge specifically for information access technology, but awareness of stylistic differences and of acceptability will be a guide to what can be expected to be found in data sources and how much effort should be put into the resolution between similar topics, into sentiment analysis, and other similar tasks.

Many timely and new texts are generated in new media and new genres with little or no editorial oversight: with new, emerging, and relatively volatile stylistic conventions; anchored into highly interactive discourse or into multimodal presentations; incorporating code switching between several languages; characterised by newly minted terms, humorous and deliberate misspellings, topic indicators (“hash tags”), and plenty of misspellings or typing errors (Karlgren 2006; Uryupina et al. 2014). This variation does not always follow the same paths across cultural and linguistic areas.

Language processing tools that are built or trained to handle standard language from e.g. news text or academic texts risk being less useful for analysis of new text. Using such tools for multi-lingual material risks skewing results across cultural areas, especially if the reader is less than fluent in the original languages.

## 2.2 *Inflection*

One of the first and most obvious differences between human languages is that of *morphology* or inflectional systems: anyone who has made the effort to learn a foreign language is familiar with the challenge of learning e.g. verb forms or plural forms, especially irregular ones. The number of different forms of a single lexical entry varies greatly between human languages. Some examples are given in Table 1. Many languages find it necessary to include information about the gender of referents (“elle est fatiguée” vs. “il est fatigué”; “śpiewał” vs. “śpiewała”); others do not. Some require tense or aspect to be marked, some do not. Some allow subjects to be omitted if understood from context (“wakarimasen”); others require subjects even when of low informational content (“es regnet”). The largest languages in the world have very spare morphology: English, Chinese, and Spanish can be analysed using very simple tools (Lovins 1968; Porter 1980). Larger languages seem to tend towards simpler morphology, and this observation has been tentatively proposed to have to do with the amount of cultural contact a larger language engages in simply through its dispersal pattern (Dahl 2004).<sup>1</sup>

---

<sup>1</sup>This would seem to be good news for language technologists with limited resources at their disposal.

**Table 1** Examples of inflectional variation given for nouns from some languages

|         |                                    | Singular     |                    | Plural   |           |
|---------|------------------------------------|--------------|--------------------|----------|-----------|
| Chinese |                                    | 虾            |                    |          |           |
| English |                                    | kipper       | kipper's           | kippers  | kippers'  |
| Swedish | Indefinite                         | sill         | sills              | sillar   | sillars   |
|         | Definite                           | sillen       | sillens            | sillarna | sillarnas |
| Finnish | ...                                | muikku       | muikun             | muikut   | muikkujen |
|         | Ablative + "not even"              | muikultakaan |                    | ...      | ...       |
|         | Adessive + our + "also" + Emphatic | ...          | muikuillannekinhan |          |           |

Chinese nouns do not inflect. English inflects less than Swedish. Finnish has thousands of possible forms for each nouns

The majority of the world's languages, if not the majority of speakers, have more elaborate morphology. Morphological analysis tools of various levels of sophistication have been developed for languages, often inspired by languages with richer morphological variation than English. These tools have been applied to various tasks such as writing aids, translation, speech recognition, and lately included as a matter of course in many information access systems.

Nouns are in most languages inflected by *number*, to distinguish between one, many, and in some cases pairs of items. In most languages nouns are also inflected by *case*, to indicate the noun's role with respect to other words in a clause. English uses the genitive form to indicate ownership; Latin uses different cases for object and various adverbial functions; Russian adds yet another case to indicate an instrument; Finnish and related languages have a dozen or so cases to indicate various positional and functional roles of nouns. Some languages indicate *definiteness* by inflection (which in English is marked by separate determiners such as *the* or *a*). Verbs in most languages carry information of a temporal and aspectual character of the event, state, or process the clause refers to. In general, adjectives exhibit less complex inflection patterns than do nouns; verbs tend to be more elaborate than nouns.

This variation directly impacts information retrieval performance. If surface variation of terms is reduced through some procedure, the recall of a retrieval system is increased—at some cost to precision—through the system retrieving documents which contain some term in a different surface form than that presented by the user in a query: if a system knows enough to find texts mentioning "festival" when a user searches for texts on "festivals" it will most likely make its users happier (Lowe et al. 1973; Lennon et al. 1981). The process where different forms of a word are collated is variously called *normalisation*, *lemmatisation*, *stemming*, or even *truncation*, depending on which engineering approach is taken to the task.

This variation in morphological systems across languages from the perspective of information access has been addressed in previous literature by e.g. Pirkola (2001) who has formulated a description of languages of the world using two variables, *index of synthesis* and *index of fusion* and examined how those variables could

be used to inform the design of practical tools for both mono- and cross-lingual information retrieval research and system development.

For English, for a long time, it was taken to be proven that normalisation by and large would not help retrieval performance (Harman 1991). Once the attention of the field moved to languages other than English, it was found that for other languages there were obvious gains to be found (Popović and Willett 1992), with the cost and utility of analysis varying across languages and across approaches as to how it is deployed (Kettunen and Airio 2006; Kettunen et al. 2007; Karlgren et al. 2008; Kettunen 2009, 2014; McNamee et al. 2009).

Not every morphological form is worth normalising. Languages such as Finnish or Basque, e.g., have several thousand theoretically possible forms for each noun. In practice only a small fraction of them actually show up in text. Taking care of the more frequent forms has clear effects on retrieval performance; other forms are more marginal, or may even reduce performance for topical retrieval tasks, if variants which make topically relevant distinctions are conflated.

Today morphological analysis components to normalise terms from text and queries, using a stem or a lemma form instead of the surface form, are used in retrieval systems as a matter of course. For some languages and some tasks, fairly simple truncation-based methods (Porter 1980, 2001) or n-gram indexing (Kamps et al. 2004) yield quite representative results, but more informed approaches are necessary for the systematic treatment of e.g. languages where affixation can include prefixes or infixes. Most systems today incorporate morphological normalisation by default for some of the larger languages and tools for the introduction of such techniques for languages with less existing technology support.

### 2.3 *Derivation and Compounding*

Derivation, the creation of new words by modifying others, and compounding, the creation of new words by combining previously known ones, are productive processes in all human languages. There is no limit to creating new words, but there is a limit in how and to what extent they can be and are included in for example translation dictionaries used in multi- or cross-language information access technology.

Derivational morphology describes how new words can be created through the use of affixes (prefixes and suffixes) combined to a word stem, e.g., *build*—*builder*—*building*. Derivation thus affects the part-of-speech and meaning of the word *build* (Akmajian et al. 1995).

Compounds can be closed (such as *classroom*), open (such as *ice cream*), or hyphenated (such as *well-being*). Human languages vary as to how they orthographically construct compounds: German, Dutch, Finnish and Swedish, e.g., favour closed compounds; English orthography is less consistent, but uses open compounds to a much greater extent. The orthographic specification is important in cross-lingual retrieval and is also related to the translation and identification of compounds

as phrases in for example English. Closed compounds are easier to handle in information access technology and in cross-language applications because there is no need for a specific identification of a “phrase” as in open compounds (Lieber and Štekauer 2009).

Splitting compounds into their constituents may be expedient for the purposes of information retrieval: the compounds may be too specific and splitting them would yield useful and content-bearing constituents, thus increasing recall of an index. This is especially true in a scenario where queries are translated from one language to another (Hedlund et al. 2001).

Doing this is not straightforward, however. A compound may be *compositional*, where the meaning of the compound is a function of some sort of its constituents, or *non-compositional* where the meaning of the constituents is non-relevant or marginally relevant to understanding the compound. Where a compound is compositional, the relation between its constituents may be difficult to predict without world knowledge: most compounds in frequent use have been lexicalised as words in their own right to some extent. In practice, frequently only some or even none of the constituents of a compound are topically relevant (such as in *strawberry*, *Erdbeere*, *fireworks*, or *windjammer*). A compound may also have several possible splits, with typically only one of them being correct (such as in *sunflower*). In languages which make free use of closed compounds these challenges are exacerbated: the Swedish *domstol* (*court of law*) can be split into *dom* and *stol*, the former being *judgment* but also the homograph personal pronoun *they* which trumps the relevant reading by frequency; the latter being *chair*, which is irrelevant; the Swedish 3-way compound *riksdagshus* (*parliament building*) can be reconstructed into *riks*, *dag* and *hus* (*realm*, *day*, *building*) which is less useful than the 2-way split into *riksdag* and *hus*, (*parliament* and *building*).AQPlease check the spelling of the term “exarcebated” in the sentence “In languages which make...”, and correct if necessary.

Many languages make use of *fogemorphemes*, glue components between information bearing constituents, for example, *-ens-* in *Herz-ens-brecher*, the German word for *heart breaker*. Handling these correctly impacts performance noticeably (Hedlund 2002; Kamps et al. 2004).

Challenges such as these make the application of compound splitting somewhat more difficult than the seemingly simple process the term itself invites (Chen 2001, 2002; Hedlund et al. 2000; Hedlund 2002; Cöster et al. 2003; Karlgren 2005).

In summary, some of the challenges with using constituents from compound splitting are that they may not express a concept similar to that expressed by the compound; may be ambiguous; may not always even be valid words.

## 2.4 Word Order and Syntactic Variation

Languages vary greatly in how strictly rule-bound the word order of their utterances is, and what that rule order is. In clauses, many languages with strict rule order such

as English, require a subject-verb-object order (Example (1-b)) in typical clauses; most languages of the world prefer subject-object-verb order (Example (1-a)) instead, and many languages use verb-subject-object (Example (1-c)). The other three orderings are quite rare in comparison. Languages with comparatively free word order still invariably exhibit a preference for a standard word order which is used when there is no reason to diverge from it, e.g. for reasons of topical emphasis.

- (1) a. *Caesar aleas amat.*  
*Caesar dice loves (Latin)*
- b. *The slow fox caught the early worm.*
- c. *Phóg an fear an muc.*  
*Kissed the man the pig. (Irish)*

With respect to single constituents, languages vary in how they organise a head word and its attributes. Adjectives can precede (Examples (2-b) and (2-c)) the noun they modify or come after (Example (2-a)); a language may prefer prepositions to postpositions.

- (2) a. *Un vin blanc sec*  
*A wine white dry (French)*
- b. *An unsurprising sample*
- c. *Bar mleczny w Częstochowie*  
*Bar milk in Czestochowa (Polish)*
- d. *A hegedű a zongora mögött van*  
*The violin the piano behind is (Hungarian)*

For any information based on more elaborate analyses than bags of words, these variations will impact the results. If e.g. a system automatically recognises multi-word phrases, word order will make a difference; if the tasks move beyond information retrieval to e.g. information extraction, sentiment analysis or other tasks, where more than word counts are instrumental to the analysis, an analysis step to identify head with respect to attribute will be necessary.

## 2.5 *Ellipsis and Anaphora*

Elliptic references in human language include omission of words that are obviously understood, but must be added to make the construction grammatically complete.

Human language users avoid repetition of referents, replacing something known by a pronoun, and sometimes omit the referent entirely. The ways in which this is done vary somewhat over languages and genres. Samples (3) are in English, with omitted bits in square brackets.

- (3) a. *Kal does not have a dog but Ari does [have a dog]*  
 b. *I like Brand A a lot. But on the whole, Brand B is better [than Brand A].*  
 c. *Bertram makes deep-V hulls. It [Bertram] takes sea really well.*

Elliptic references are challenging from the point of view of information retrieval, because search words may be omitted in the text (Pirkola and Järvelin 1996). Such omissions will impact retrieval efficiency in that the relative frequencies of terms implicitly understood by the author and reader of a text may be under-represented by an indexing tool. This effect is likely to be marginal, but more importantly, analyses and tasks with more semantic sophistication, which depend on associating a feature or characteristic with some referent will be difficult unless the referent in question is explicitly mentioned. Sentiment analysis (Steinberger et al. 2011) and keyword proximity based retrieval (Pirkola and Järvelin 1996) are examples.

## 2.6 Digitisation of Collections and Historical Depth

When originally non-digital material, such as old newspapers and books, are digitized, the process starts with the documents scanned into image files. From these image files one needs to sort out texts and possible non-textual data, such as photographs and other pictorial representations. Texts are recognized from the scanned pages with Optical Character Recognition (OCR) software. OCR for modern text types and fonts is considered to be a solved problem that yields high quality results, but results of historical document OCR are still far from that level (Piotrowski 2012). Most recently, Springmann and Lüdeling (2017) report high word-level recognition accuracies (ranging from 76% to 97%) based on applying trainable Neural Network-based OCR to a diachronic corpus of scanned images of books printed between 1478 and 1870. This type of corpus is especially demanding for OCR due to many types of variation present in the manuscripts—including linguistic changes (e.g., spelling, word formation, word order) and extra-linguistic changes (e.g., medium, layout, scripts, and technology).

Digitization of old books, newspapers and other material has been an on-going effort for more than 20 years in Europe. Its results can be seen e.g. in large multilingual newspaper collections, such as Europeana (<http://www.europeana-newspapers.eu/>). Europeana contains 18 million pages in 16 languages (Pletschacher et al. 2015). Scandinavian countries, e.g., have available over 80 million pages of digitized historical newspapers (Pääkkönen et al. 2018). Single newspaper archives,

such as Times of London 1785–2012, or La Stampa 1867–2005, can already contain several million or over 10 million pages.

Europeana has estimated word level quality of its contents. For most of the included major languages, word correctness rate is about 80% or slightly more, but for Finnish, Old German, Latvian, Russian, Ukrainian and Yiddish, correctness rates are below 70% (Pletschacher et al. 2015). Thus smaller languages and content published in more complicated scripts may have a disadvantage in their quality.

OCR errors in the digitized newspapers and journals may impact collection quality. Poor OCR quality obviously renders documents from the collections less readable and comprehensible for human readers but also less amenable to on-line search and further natural language processing or analysis (Taghva et al. 1996; Lopresti 2009). Savoy and Naji (2011), for example, showed how retrieval performance decreases with OCR error corrupted documents quite severely.

The same level of retrieval quality decrease is shown in results from the confusion track at TREC 5 (Kantor and Voorhees 2000). The end result effect of OCR errors is not clear cut, however. Tanner et al. (2009) suggest that word accuracy rates less than 80% are harmful for search, but when the word accuracy is over 80%, fuzzy search capabilities of search engines should manage the problems caused by word errors. The probabilistic model developed by Mittendorf and Schäuble (2000) for data corruption seems to support this, at least for longer documents and longer queries. Empirical results by Järvelin et al. (2016) on a Finnish historical newspaper search collection show that fuzzy matching will help only to a limited degree if the collection is of low quality.

One aspect of retrieval performance of poor OCR quality is its effect on ranking of the documents (Mittendorf and Schäuble 2000): badly OCRed documents may be quite low in the result list if they are found at all. In practice these kinds of drops in retrieval and ranking performance mean that the user will lose relevant documents: either they are not found at all by the search engine or the documents are so low in the ranking list that the user may never reach them while browsing the result list. Some examples of this in the work of digital humanities scholars are discussed e.g. by Traub et al. (2015).

Correcting OCR errors in a historical corpus can be done at access time or at indexing time by filtering index terms through authoritative lexical resources, pooling the output from several OCR systems (under the assumption they make different errors) or using distributional models to find equivalents for unknown words. These are all methods tested and used for OCR correction. As observed by Volk et al. (2011), built-in lexicons of commercial OCR systems do not cover nineteenth century spelling, dialectal or regional spelling variants, or proper names of e.g. news material from previous historical eras. Afli et al. (2016) propose that statistical machine translation can be a beneficial method for performing post-OCR error correction for historical French.



### 3 Reliance on Resources

Languages with few developed language technology resources are sometimes called *low-density languages*. While the concept is somewhat vague, it can be useful in as much as it makes clear that languages with a small number of speakers may be well served by language technology, whereas widely used languages may or may not be considered low-density. Examples of early studies in African low-density languages in cross-language information retrieval include Cosijn et al. (2004) (Afrikaans-English) and Argaw et al. (2004, 2005), Argaw and Asker (2006), Argaw (2007) (Amharic-English). Both explored the effectiveness of query translation utilizing topic (source) word normalization, bilingual dictionary-lookup, and removal of stop words as process components. The first study reports the development of a simplified Afrikaans normalizer; the latter used semi-automatic Amharic stemming (prefix and suffix stripping).

#### 3.1 Dictionaries and Lexical Resources

Various types of lexical resources are necessary in Natural Language Processing (NLP). Monolingual dictionaries are used in morphological analysis for producing lemmas, and for decomposing compound words—and as the necessary step for subsequent phases in NLP, e.g., for recognising noun phrases or names; for recognising the target of some expressed attitude; or for extracting emerging topics from a stream of text. Not least in translating queries or other specifications of information needs, dictionaries will form a crucial component (Pirkola et al. 2001; Hedlund et al. 2004).

Synonym dictionaries or thesauri are used for expanding queries, to add recall to a narrowly posed information need. Bilingual dictionaries may be intended either for human readers (and thus contain verbose definitions) or alternatively intended for automatic translation components (transfer dictionaries) either for text translation or for e.g. query translation. It is a non-trivial problem to transfer a bilingual dictionary intended for humans into a transfer dictionary (Hull and Grefenstette 1996).

#### 3.2 Automatic Machine Translation of Queries and the Challenge of Out-of-Vocabulary Terms

Over the years at CLEF and elsewhere, many researchers have performed and continuously perform experiments to use existing automatic and semi-automatic machine translation resources to translate queries. Various technologies have been tested against each other, against manual human translation, against translated indexes, or against translated target documents (Airio 2008). The quality of retrieval results, noted by practically all such studies, depends on two factors. Firstly, that

publicly available translation resources are primarily intended to provide a *crude* translation designed for human readers, not a *raw* translation for continued editing or use in further processes such as retrieval (Karlgren 1981). Translations by web resources tend to resolve ambiguities with this in mind, and thus occasionally reducing information present in the original query. This can be ameliorated by systems that use other lexical resources to enrich the translated query (Herbert et al. 2011; Leveling et al. 2009; Saleh and Pecina 2016).

Secondly, and more obviously, coverage of the translation resource. Out Of Vocabulary (OOV) words, i.e., words not found in translation dictionaries, are the major challenge for CLIR, machine translation, and other multilingual language processing tasks and information systems where translation is part of the system. In particular in scientific and technical domains OOV words are often keywords in texts, and if the system is unable to translate the most important words its effectiveness may substantially decrease. Proper names form another word category causing translation problems: while they should not be translated in principle, their surface forms in different languages may differ due to transliteration and inflection. The tools to handle OOV translation include: approximate string matching (fuzzy matching) through methods such as Soundex, character-level n-grams (skip-grams), and edit distance; reverse transliteration e.g. as in Transformation rule based translation in which a word in one language (e.g. Finnish *somatologia*  $\longleftrightarrow$  English *somatology* or Finnish *Tsetsenia*  $\longleftrightarrow$  English *Chechnya*) based on the regular correspondences between the characters in spelling variants (Pirkola et al. 2003, 2007; Toivonen et al. 2005).

## 4 Challenges

The challenges entailed by cross-linguistic variation can be summarized to be about resources: lack of them, cost of acquiring and maintaining them, and low utility of seemingly relevant tools developed e.g. by computational linguists. Tools built by computational linguists do not always improve results on large scale information processing tasks, since they are built for a different purpose than information access (Table 2).

While the field of information access research has human communicative behaviour as its main object of study and processing texts and other human communicative expressions to understand their content, linguistic theory has as its goals to explain the structure and regularities of human language. These goals are related but are not perfectly aligned. Obviously linguists would do well to validate their theories by application to information access, but they lack an understanding of what needs are prioritised; information access researchers must formulate requirements for better analyses for computationally oriented linguists to work on, and these requirements need to be formulated at an operationally adequate level of abstraction. These discussions and analyses are what CLEF and other related

**Table 2** Challenges in utilizing various resources in information access

| Resource or technology                | Monolingual Information Retrieval  | Crosslingual and Multilingual Information Retrieval  |
|---------------------------------------|--|--|
| Lexicons or translation dictionaries  | <ul style="list-style-type: none"> <li>– need to create resources per se (especially in low-density languages)</li> <li>vocabulary issues:               <ul style="list-style-type: none"> <li>– insufficient coverage</li> <li>– domain-specific needs (e.g., social media, historical texts, etc.)</li> <li>– OOV words (e.g., proper names)</li> <li>– control and cost of updating</li> </ul> </li> </ul> | <ul style="list-style-type: none"> <li>– need to create transfer dictionaries appropriate for CLIR</li> <li>vocabulary issues:               <ul style="list-style-type: none"> <li>– insufficient coverage</li> <li>– excessive number of translations</li> <li>– domain-specific needs</li> <li>– OOV words</li> <li>– control of updating the vocabularies</li> <li>– cost of updating</li> </ul> </li> </ul> |
| Stop word lists                       | <ul style="list-style-type: none"> <li>– need to create and tune stop word vocabularies for the particular application and domain</li> </ul>   | <ul style="list-style-type: none"> <li>– same as in monolingual case (but for both source and target languages)</li> </ul>   |
| Normalising and lemmatisation methods | <ul style="list-style-type: none"> <li>– vocabulary issues (in lemmatisation)</li> <li>– understemming, overstemming, and incorrect processing (in stemming)</li> <li>– linguistically correct processing may be inappropriate from the point of view of IR (generation of nonsense words)</li> </ul>  | <ul style="list-style-type: none"> <li>– vocabulary issues (coverage, updating, etc.)</li> <li>– need to detect and translate multi-word phrases (in phrase-oriented languages)</li> <li>– need to decompound and translate compound words written together (in compound-oriented languages)</li> </ul>  |
| Fuzzy string matching                 | <ul style="list-style-type: none"> <li>– applicability may be language-specific</li> <li>– effectiveness and efficiency issues</li> </ul>  | <ul style="list-style-type: none"> <li>– applicability may be language pair-specific</li> <li>– effectiveness and efficiency issues</li> </ul>   |
| Generative methods                    | <ul style="list-style-type: none"> <li>– need to design and implement the method (in low-density languages)</li> <li>– relatively high number of potential candidate words created (in highly inflectional languages)</li> <li>– efficiency issues</li> <li>– challenges of special domains (e.g., creating expressions matching noisy OCR text)</li> </ul>  | <ul style="list-style-type: none"> <li>– same as in monolingual case</li> <li>– here the idea is to generate query expansions for the target language in which normalization or lemmatization may not be available or appropriate (e.g., in web domain)</li> </ul>   |
| Comparable corpora                    | (not applicable)   | <ul style="list-style-type: none"> <li>– availability of appropriate corpora for the particular language-pairs in need</li> <li>– appropriateness of the alignment methods</li> </ul>  |

forums are for; the output could be communicated in clearer terms, in the form of clearly formulated usage scenarios or use cases, for further discussions with application-minded linguists.

## References

- Afli H, Qiu Z, Way A, Sheridan P (2016) Using SMT for OCR error correction of historical texts. In: 10th international conference on language resources and evaluation, LREC. European Language Resources Association, France, pp 962–966
- Airio E (2008) Who benefits from CLIR in web retrieval? *J Doc* 64(5):760–778
- Akmajian A, Demers R, Farmer A, Harnish R (1995) *Linguistics: an introduction to language and communication*, 4th edn. MIT Press, Cambridge

- Argaw AA (2007) Amharic-English information retrieval with pseudo relevance feedback. In: Nardi A, Peters C, Ferro N (eds) CLEF 2007 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1173/>
- Argaw AA, Asker L (2006) Amharic-English information retrieval. In: Workshop of the cross-language evaluation forum for European languages. Springer, Berlin, pp 43–50
- Argaw AA, Asker L, Cöster R, Karlgren J (2004) Dictionary-based Amharic–English information retrieval. In: Workshop of the cross-language evaluation forum for European languages. Springer, Berlin, pp 143–149
- Argaw AA, Asker L, Cöster R, Karlgren J, Sahlgren M (2005) Dictionary-based Amharic-French information retrieval. In: Workshop of the cross-language evaluation forum for European languages. Springer, Berlin, pp 83–92
- Chen A (2001) Multilingual information retrieval using English and Chinese queries. In: Workshop of the cross-language evaluation forum for European languages. Springer, Berlin, pp 44–58
- Chen A (2002) Cross-language retrieval experiments at clef 2002. In: Workshop of the cross-language evaluation forum for European languages, Springer, Berlin, pp 28–48
- Cosijn E, Keskustalo H, Pirkola A, De Wet K (2004) Afrikaans-English cross-language information retrieval. In: Bothma T, Kaniki A (eds) Proceedings of the 3rd biennial DISSAnet conference, Pretoria, pp 97–100
- Cöster R, Sahlgren M, Karlgren J (2003) Selective compound splitting of Swedish queries for boolean combinations of truncated terms. In: Workshop of the cross-language evaluation forum for European languages. Springer, Berlin, pp 337–344
- Dahl Ö (2004) The growth and maintenance of linguistic complexity, vol 71. John Benjamins, Amsterdam
- Dryer MS, Haspelmath M (2011) The world atlas of language structures online. Max Planck Digital Library, München. <http://wals.info>
- Gollins T, Sanderson M (2001) Improving cross language retrieval with triangulated translation. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 90–95
- Harman D (1991) How effective is suffixing? *J Am Soc Inf Sci* 42:7–15
- Hedlund T (2002) Compounds in dictionary-based cross-language information retrieval. *Inf Res* 7(2):7-2
- Hedlund T, Keskustalo H, Pirkola A, Sepponen M, Järvelin K (2000) Bilingual tests with Swedish, Finnish, and German queries: dealing with morphology, compound words, and query structure. In: Workshop of the cross-language evaluation forum for European languages, Springer, Berlin, pp 210–223
- Hedlund T, Pirkola A, Järvelin K (2001) Aspects of Swedish morphology and semantics from the perspective of mono-and cross-language information retrieval. *Inf Process Manage* 37(1):147–161
- Hedlund T, Airio E, Keskustalo H, Lehtokangas R, Pirkola A, Järvelin K (2004) Dictionary-based cross-language information retrieval: learning experiences from clef 2000–2002. *Inf Retrieval* 7(1–2):99–119
- Herbert B, Szarvas G, Gurevych I (2011) Combining query translation techniques to improve cross-language information retrieval. In: Proceedings of the 33D European conference on information retrieval. Springer, Berlin
- Hull DA, Grefenstette G (1996) Querying across languages: a dictionary-based approach to multilingual information retrieval. In: Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 49–57
- Järvelin A, Keskustalo H, Sormunen E, Saastamoinen M, Kettunen K (2016) Information retrieval from historical newspaper collections in highly inflectional languages: A query expansion approach. *J Assoc Inf Sci Technol* 67(12):2928–2946
- Kamps J, Monz C, De Rijke M, Sigurbjörnsson B (2004) Language-dependent and language-independent approaches to cross-lingual text retrieval. In: Peters C, Braschler M, Gonzalo J, Kluck M (eds) Comparative evaluation of multilingual information access systems: fourth

- workshop of the cross-language evaluation forum (CLEF 2003) revised selected papers. Lecture notes in computer science (LNCS), vol 3237. Springer, Heidelberg
- Kantor PB, Voorhees EM (2000) The TREC-5 confusion track: comparing retrieval methods for scanned text. *Inf Retrieval* 2(2):165–176
- Karlgren H (1981) Computer aids in translation. *Stud Linguist* 35(1–2):86–101
- Karlgren J (2005) Compound terms and their constituent elements in information retrieval. In: Proceedings of the 15th Nordic conference of computational linguistics (NoDaLiDa). University of Joensuu, Finland, pp 111–115
- Karlgren J (ed) (2006) New text—wikis and blogs and other dynamic text sources. In: Proceedings of the EACL06 workshop. European Chapter of the Association for Computational Linguistics
- Karlgren J, Dalianis H, Jongejan B (2008) Experiments to investigate the connection between case distribution and topical relevance of search terms. In: 6th international conference on language resources and evaluation, LREC
- Kettunen K (2009) Reductive and generative approaches to management of morphological variation of keywords in monolingual information retrieval: an overview. *J Doc* 65(2):267–290
- Kettunen K (2014) Can type-token ratio be used to show morphological complexity of languages? *J Quant Linguist* 21(3):223–245
- Kettunen K, Airio E (2006) Is a morphologically complex language really that complex in full-text retrieval? In: Advances in natural language processing. Springer, Berlin, pp 411–422
- Kettunen K, Airio E, Järvelin K (2007) Restricted inflectional form generation in morphological keyword variation. *Inf Retrieval* 10(4–5):415–444
- Lehtokangas R, Airio E, Järvelin K (2004) Transitive dictionary translation challenges direct dictionary translation in *clir*. *Inf Process Manage* 40(6):973–988
- Lennon M, Peirce DS, Tarry BD, Willett P (1981) An evaluation of some conflation algorithms for information retrieval. *Information Scientist* 3(4):177–183
- Leveling J, Zhou D, Jones GJF, Wade V (2009) TCD-DCU at TEL@CLEF 2009: document expansion, query translation and language modeling. In: Borri F, Nardi A, Peters C, Ferro N (eds) CLEF 2009 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613–0073. <http://ceur-ws.org/Vol-1175/>
- Lewis MP, Simons GF, Fennig CD, et al (2009) *Ethnologue: languages of the world*, vol 16. SIL international, Dallas. <http://www.ethnologue.com>
- Lieber R, Štekauer P (2009) *The Oxford handbook of compounding*. Oxford University Press, Oxford
- Lopresti D (2009) Optical character recognition errors and their effects on natural language processing. *Int J Doc Anal Recogn* 12(3):141–151
- Lovins JB (1968) Development of a stemming algorithm. MIT Information Processing Group, Electronic Systems Laboratory, Cambridge
- Lowe TC, Roberts DC, Kurtz P (1973) Additional text processing for on-line retrieval (the radcol system), vol 1. Tech. rep., DTIC Document
- McNamee P, Nicholas C, Mayfield J (2009) Addressing morphological variation in alphabetic languages. In: Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 75–82
- Mittendorf E, Schäuble P (2000) Information retrieval can cope with many errors. *Inf Retrieval* 3(3):189–216
- Pääkkönen T, Kettunen K, Kervinen J (2018) Digitisation and digital library presentation system—a resource-conscientious approach. In: Proceedings of 3D conference on digital humanities in the Nordic countries, CEUR-WS.org, pp 297–305
- Piotrowski M (2012) Natural language processing for historical texts. *Synth Lect Hum Lang Technol* 5(2):1–157
- Pirkola A (2001) Morphological typology of languages for IR. *J Doc* 57(3):330–348
- Pirkola A, Järvelin K (1996) The effect of anaphor and ellipsis resolution on proximity searching in a text database. *Inf Process Manage* 32(2):199–216
- Pirkola A, Hedlund T, Keskustalo H, Järvelin K (2001) Dictionary-based cross-language information retrieval: problems, methods, and research findings. *Inf Retrieval* 4(3–4):209–230

- Pirkola A, Toivonen J, Keskustalo H, Visala K, Järvelin K (2003) Fuzzy translation of cross-lingual spelling variants. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 345–352
- Pirkola A, Toivonen J, Keskustalo H, Järvelin K (2007) Frequency-based identification of correct translation equivalents (FITE) obtained through transformation rules. *ACM Trans Inf Sys* 26(1):2
- Pletschacher S, Clausner C, Antonacopoulos A (2015) European newspapers OCR workflow evaluation. In: Proceedings of the 3rd international workshop on historical document imaging and processing. ACM, New York, pp 39–46
- Popović M, Willett P (1992) The effectiveness of stemming for natural-language access to slovene textual data. *J Am Soc Inf Sci* 43(5):384–390
- Porter MF (1980) An algorithm for suffix stripping. *Program* 14(3):130–137
- Porter MF (2001) Snowball: a language for stemming algorithms
- Rehm G, Uszkoreit H (2012) Meta-net white paper series: Europe’s languages in the digital age
- Saleh S, Pecina P (2016) Reranking Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval. In: Fuhr N, Quaresma P, Gonçalves T, Larsen B, Balog K, Macdonald C, Cappellato L, Ferro N (eds) Experimental IR meets multilinguality, multimodality, and interaction. Proceedings of the seventh international conference of the CLEF association (CLEF 2016). Lecture notes in computer science (LNCS), vol 9822, Springer, Heidelberg, pp 54–68
- Savoy J, Naji N (2011) Comparative information retrieval evaluation for scanned documents. In: Proceedings of the 15th WSEAS international conference on Computers, pp 527–534
- Springmann U, Lüdeling A (2017) OCR of historical printings with an application to building diachronic corpora: a case study using the RIDGES corpus. *Digit Humanit* Q11(2)
- Steinberger J, Lenkova P, Kabadjov MA, Steinberger R, Van der Goot E (2011) Multilingual entity-centered sentiment analysis evaluated by parallel corpora. In: Recent advances in natural language processing, pp 770–775
- Taghva K, Borsack J, Condit A (1996) Evaluation of model-based retrieval effectiveness with OCR text. *ACM Trans Inf Syst* 14(1):64–93
- Tanner S, Muñoz T, Ros PH (2009) Measuring mass text digitization quality and usefulness. lessons learned from assessing the OCR accuracy of the British library’s 19th century online newspaper archive. *D-lib Mag* 15(7/8):1082–9873
- Toivonen J, Pirkola A, Keskustalo H, Visala K, Järvelin K (2005) Translating cross-lingual spelling variants using transformation rules. *Inf Process Manag* 41(4):859–872
- Traub MC, van Ossenbruggen J, Hardman L (2015) Impact analysis of OCR quality on research tasks in digital archives. In: Kapidakis S, Mazurek C, Werla M (eds) International conference on theory and practice of digital libraries. Lecture notes in computer science (LNCS), vol 9316. Springer, Heidelberg, pp 252–263
- Uryupina O, Plank B, Severyn A, Rotondi A, Moschitti A (2014) Sentube: a corpus for sentiment analysis on youtube social media. In: 9th international conference on language resources and evaluation, LREC
- Velupillai V (2012) An introduction to linguistic typology. John Benjamins Publishing, Amsterdam
- Volk M, Furrer L, Sennrich R (2011) Strategies for reducing and correcting OCR errors. *Language technology for cultural heritage*, pp 3–22

# Multi-Lingual Retrieval of Pictures in ImageCLEF



Paul Clough and Theodora Tsikrika

**Abstract** CLEF first launched a multi-lingual visual information retrieval task in 2003 as part of the ImageCLEF track. Several such tasks subsequently followed, encompassing both the medical and non-medical domains. The main aim of such ad hoc image retrieval tasks was to investigate the effectiveness of retrieval approaches that exploit textual and visual evidence in the context of large and heterogeneous collections of images that are searched for by users with diverse information needs. This chapter presents an overview of the image retrieval activities within ImageCLEF from 2003 to 2011, focusing on the non-medical domain and, in particular, on the photographic retrieval and Wikipedia image retrieval tasks. We review the available test collections built in the context of these activities, present the main evaluation results, and summarise the contributions and lessons learned.

## 1 Introduction

Visual information indexing and retrieval has been an active research area since the 1990s and the subject of much research effort (Del Bimbo 1999). Visual information can include sketches, photographs, 3D images, videos, etc., and low-level descriptors such as texture, shape and colour could be used for indexing and retrieval. In addition, visual media can be accompanied by textual metadata, such as date and producer (*content-independent* metadata), as well as descriptions of visual content or assigned keywords representing high-level concepts (*content-dependent* and *descriptive* metadata). In this chapter, we focus on the retrieval

---

P. Clough (✉)  
Information School, University of Sheffield, Sheffield, UK  
e-mail: [p.d.clough@sheffield.ac.uk](mailto:p.d.clough@sheffield.ac.uk)

T. Tsikrika  
Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki,  
Greece  
e-mail: [theodora.tsikrika@iti.gr](mailto:theodora.tsikrika@iti.gr)

of 2D still images in situations where accompanying metadata may exist in multiple languages and/or users may submit their queries in multiple languages. The effective combination/fusion of features derived from the visual content and textual metadata provides opportunities to improve retrieval performance and offers one of the main areas of current interest, along with image classification and object recognition (Russakovsky et al. 2015), application of deep learning techniques (Wan et al. 2014), detection of fake or misleading multimedia content (Boididou et al. 2014), efficient and effective handling of big data (Husain and Bober 2017), and others.

Paramount to developing and improving visual information retrieval systems is their evaluation. Test collections for visual information retrieval, consisting of multimedia resources, topics, and associated relevance assessments (ground truth), enable the reproducible and comparative evaluation of different approaches, algorithms, theories, and models, through the use of standardised datasets and common evaluation methodologies. Such test collections are typically built in the context of evaluation campaigns that experimentally assess the worth and validity of new ideas in a laboratory setting within regular evaluation cycles. Over the years, several such evaluation activities have helped to foster innovation and provided the infrastructure and resources for the systematic evaluation of image and video retrieval systems, including the Internet Imaging Benchathlon,<sup>1</sup> the TREC Video Retrieval Evaluation,<sup>2</sup> the MediaEval Benchmarking Initiative for Multimedia Evaluation,<sup>3</sup> and the Cross-Language Image Retrieval (ImageCLEF)<sup>4</sup> evaluation campaigns; further evaluation resources are also available.<sup>5,6</sup>

The focus of this chapter is on the ImageCLEF initiative, which first ran in 2003. The first 10 years of the evaluation activities of this CLEF track have been summarised in the ImageCLEF book (Müller et al. 2010). The aim of the track was to investigate multi-lingual image retrieval across varying tasks and domains. Broadly speaking, the tasks fell within the following categories: ad hoc retrieval, object and concept recognition, and interactive image retrieval within the domains of medical and non-medical. The latter domain included historical archives, general photographic collections, and Wikipedia images. Re-usable evaluation resources were created to evaluate monolingual, cross-lingual, and multi-lingual retrieval systems. This chapter focuses on the non-medical image retrieval tasks within ImageCLEF and we summarise them in the following way, similar to the structure presented in Paramita and Grubinger (2010):

---

<sup>1</sup><https://sourceforge.net/projects/benchathlon/>.

<sup>2</sup><http://trecvid.nist.gov/>.

<sup>3</sup><http://www.multimediaeval.org/>.

<sup>4</sup><http://www.imageclef.org/>.

<sup>5</sup><http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm>.

<sup>6</sup><http://datasets.visionbib.com/>.



- Retrieval of images from the St Andrews University Library historic photographic archive with structured English metadata 2003–2005 (Sect. 2);
- Retrieval of images from the IAPR-TC12 photographic collection with structured multi-lingual metadata 2006–2008 (Sect. 3);
- Retrieval of images from the Belga news image collection with unstructured English textual descriptions 2009 (Sect. 4);
- Retrieval of images from Wikipedia with structured English (2008–2009) and multi-lingual (2010–2011) metadata (Sect. 5).

We then present a summary of lessons learned and main contributions of ImageCLEF for multi-lingual photographic retrieval (Sect. 6), and finally conclude this chapter (Sect. 7).

## 2 Retrieval from Photographic Archives

Cross-lingual retrieval from the St Andrews University Library (Scotland) historic photographic collection was one of the tasks of ImageCLEF 2003 (Clough and Sanderson 2004). Participants were provided with resources to evaluate systems for an ad hoc retrieval task: the query is unknown to the system in advance. Resources included a collection of 28,133 historic photographs with English captions comprising both content-independent metadata (e.g., location and date) and content-dependent metadata (e.g., description and categories) produced by librarians at St Andrews University Library. Figure 1 shows an example image and metadata from the collection which is described fully in Clough et al. (2006b). The collection provided a number of challenges, including captions and queries short in length, images of varying content and quality (i.e., mostly black and white thereby limiting the effectiveness of using colour as a visual feature), captions containing



**Fig. 1** Sample image and caption from St. Andrews University Library (Copyright St Andrews University Library)

text not directly associated with the visual content of an image (e.g., expressing something in the background), and use of colloquial and domain-specific language in the caption. All metadata was provided in English only and therefore the task was an  $X \rightarrow$  English bilingual retrieval task.

Participants were also provided with topics in English (50 in 2003, 25 in 2004 and 28 in 2005) resembling the usual TREC format consisting of a shorter *title* and longer *narrative*. In addition, the topics included example relevant images to enable the testing of different query modalities. For evaluating cross-lingual retrieval performance, the titles were manually translated into different source languages which varied throughout the years: 7 languages in 2003 (English, Italian, German, Dutch, French, Spanish and Chinese); 12 languages in 2004; and 31 in 2005 (24 used by the participating groups). Topics were selected based on analysing query logs from St Andrews University Library and subsequently modified to make them more suitable as a test query set, e.g. inclusion of query modifiers and use of visual concepts.

The retrieval tasks for 2003–2005 were: given a topic in language  $X$  (e.g., “fishermen in boat”) retrieve as many relevant images as possible from the document collection. This aimed to simulate the situation in which a user expresses their need in text in a language different from the collection and requires a visual document to fulfil their search request (e.g., searching an on-line art gallery or stock photographic collection). Relevance judgements were produced by the organisers and then used to compute measures of retrieval effectiveness, such as Precision at 10 and MAP. The task attracted 4 groups in 2003, 12 groups in 2004 and 11 (submitting) groups in 2005. Generally results showed improved retrieval performance for combining text and visual methods and cross-lingual results comparable to a monolingual English baseline. Full details of the results can be found in the track overview papers Clough and Sanderson (2004), Clough et al. (2005, 2006a) and Paramita and Grubinger (2010).

### 3 Retrieval of Images from the IAPR-TC12 Collection

In 2006, a more general collection of images from the Technical Committee 12 (TC-12) of the International Association of Pattern Recognition called the IAPR TC-12 collection was used. This contained 20,000 images accompanied by structured metadata in three languages: English, German and Spanish (see Fig. 2). The images were more general in theme than the St. Andrews University Library collection and included colour images, thereby enabling the use of colour-based visual methods. More detailed information about the IAPR-TC12 collection can be found in Grubinger et al. (2006). In 2006 and 2007 participants were provided with resources to evaluate systems for ad hoc multi-lingual retrieval. The goals of the evaluation over this period included varying the ‘completeness’ of accompanying metadata and combining textual and visual retrieval methods. In 2008 the IAPR-TC12 collection was still used but to study the notion of *diversity*—retrieve as many *different* relevant images in the top  $n$  as possible. This was a hot topic of research

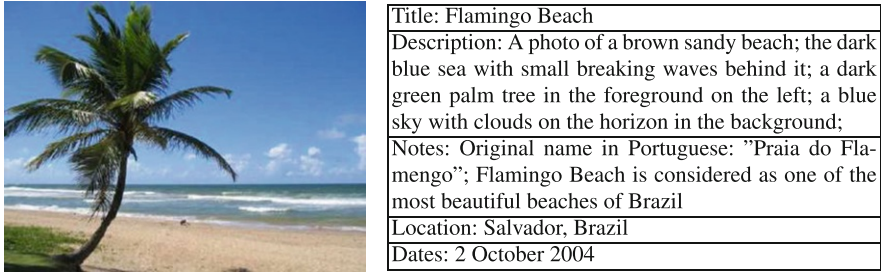


Fig. 2 Sample image and caption of the IAPR TC-12 collection

at the time (Clarke et al. 2008; Agrawal et al. 2009) and continues to attract interest (see, e.g., Santos et al. 2015).

In addition to the document collection, a set of 60 topics were developed to assess varying types of linguistic and pictorial attributes. Similar to previous years the topic consisted of a topic title (i.e., the user’s query); longer narrative description and example images to facilitate query-by-visual-example approaches. In all years, the topic titles were provided in 16 languages. However, in 2008 an additional cluster tag was added to the topic description. From the 60 topics for assessing diversity, 39 were selected in which there were obvious sub-sets (or clusters) of images relevant to the query. For example, relevant results for the query “destinations in Venezuela” could be clustered into photos from different locations within the country of Venezuela. The existing relevant images for the 39 topics were clustered, mainly based on location.

Results were evaluated for retrieval effectiveness using various measures, including MAP and Precision at 20. To evaluate diversity a measure called Cluster Recall was used, a measure that calculated the proportion of retrieved clusters to all available clusters for a particular topic (Zhai et al. 2003). The task attracted 12 groups in 2006, 20 groups in 2007 and 24 submitting in 2008. Results and analyses can be found in the track overview papers (Clough et al. 2007; Grubinger et al. 2008; Arni et al. 2009).

## 4 Retrieval of Diverse News Images from Belga

In 2009 a new collection from the Belgian news agency, Belga,<sup>7</sup> was introduced. Compared to previous years this offered a much larger document collection, comprising 498,920 photos with unstructured captions in English (see Fig. 3). The focus of the photographic retrieval task in 2009 was around evaluating retrieval systems that promote diverse results rankings. This includes ambiguous queries

<sup>7</sup><http://www.belga.be>.



837661 MOS06-20020212-MOSCOW, RUSSIAN FEDERATION: Russian Foreign Minister Igor Ivanov (L) shakes hands with Afghanistan's interim Defence Minister General Mohammad Qasim Fahim (R) to start their talks at the Foreign Ministry in Moscow on Tuesday 12 February 2002. Russia will give the technical and logistical assistance to Afghanistan's army but will not train Afghan military specialists, it was announced. EPA PHOTO EPA-SERGEI CHIRIKOV-vk-fob

**Fig. 3** Sample image and caption from Belga (Copyright Belga News Agency)

with multiple distinct meanings (e.g., Java the island; Java the drink; Java the programming language) and queries referring to broad topics that have multiple relevant aspects associated with them (e.g., London weather; London tourist information; London history, etc.).

To develop queries for the 2009 track, Belga provided a list of 1,402,990 queries submitted to the Belga search engine from 1 January 2008 to 31 December 2008. An approach was developed to identify potential query topics and sub-topics based on analysing query reformulations (Paramita et al. 2009a). For example, the query “Beckham” is often refined with examples such as “David Beckham”, “Victoria Beckham”, “Brooklyn Beckham” etc. This offers potential aspects of a query that can simulate clusters against which to evaluate diversity. The organisers produced 50 queries, with many broad and under-specified (e.g., “Belgium”), and others being highly ambiguous (e.g., “Prince” and “Euro”). Of the 50 queries, 25 were randomly selected to be released with information including the title, cluster title, cluster description and example image. The remaining 25 queries contained no information about the kind of diversity expected and simply gave visual exemplars. Precision at 10 and Cluster Recall were used to evaluate submissions.

In total, 19 groups submitted to the photographic retrieval task for ImageCLEF in 2009. Results showed that participants were able to generate runs of high diversity and relevance. Findings showed that submissions based on using mixed modalities (i.e., combinations of visual and textual features) performed best compared to those using only text-based methods or content-based image retrieval alone (Paramita et al. 2009b).

## 5 Retrieval of Wikipedia Images

Cross-lingual image retrieval from Wikipedia was added in 2008 with the goal to simulate large-scale image retrieval in realistic settings, such as the Web, where available images cover highly diverse subjects, have highly varied visual properties, and might include noisy, user-generated textual descriptions of varying lengths and quality. In light of this, the Wikipedia collaborative encyclopaedia is a suitable data source. The Wikipedia image retrieval task was actually first set up in 2006 as part of the activities of the INEX Multimedia track (Westerveld and van Zwol 2007),

but moved to ImageCLEF in 2008, which formed a more natural environment for hosting this type of benchmark and also attracted more participants from the content-based image retrieval community. The task ran until 2011 with the main aim to support ad hoc image retrieval evaluation using large-scale collections of Wikipedia images and their user-generated annotations, and to investigate the effectiveness of multimodal image retrieval approaches that combine textual and visual features.

During the four years the task ran as part of ImageCLEF (Tsirikika and Kludas 2009, 2010; Popescu et al. 2010; Tsirikika et al. 2011b), two image collections were used: (1) the Wikipedia INEX Multimedia collection (Westerveld and van Zwol 2007) in 2008 and 2009 (i.e., a cleaned-up version of the collection that had been built in the context of the INEX activities), and (2) the Wikipedia Retrieval 2010 collection (Popescu et al. 2010) in 2010 and 2011. All content selected for inclusion in the collections was licensed under Creative Commons to facilitate distribution, provided that the original license terms were respected.

The Wikipedia INEX Multimedia collection contained 151,519 images and associated textual annotations extracted from the English Wikipedia. The Wikipedia Retrieval 2010 collection consisted of 237,434 images selected to cover similar topics in English, German, and French. Similarity of coverage across the different languages was maintained by retaining images embedded in articles with versions in all three languages and at least one image in each version. The main differences between the two collections include the latter being almost 60% larger than the former and its images accompanied by (1) annotations in multiple languages and (2) links to the article(s) that contained the image. This helped to reproduce the conditions of Web image search, where images are often embedded within Web pages with long textual descriptions. Figures 4 and 5 show examples of the images in the two collections, respectively, and their associated textual annotations.

Topics representing diverse multimedia information needs were developed: task participants were provided with topics consisting of the topic title and image examples; while, similar to TREC and the other photographic retrieval tasks



**Fig. 4** Sample image and its associated user-generated annotation in English from the Wikipedia INEX multimedia collection



**Fig. 5** Sample image and its associated user-generated annotations in the three languages from the ImageCLEF 2010 Wikipedia image collection

described in this chapter, the assessors were also provided with an unambiguous description (narrative) of the type of relevant and irrelevant results. There were 75 topics in 2008, 45 in 2009, 70 in 2010, and 50 in 2011. Topic creation was collaborative in 2008 and 2009, with the participants proposing topics from which the organisers selected a final list; participation in topic creation was mandatory in 2008 and optional in 2009. In 2010 and 2011, the task organisers selected the topics after performing a statistical analysis of image search engine queries logged by the Belga News Agency image search portal in 2010 and by Exalead<sup>8</sup> in 2011. Mean topic length varied between 2.64 and 3.10 words per topic (similar to standard Web image search queries).

Following the structure of the collection, only English topics were provided in 2008 and 2009; German and French translations of the English topics were also provided in 2010 and 2011. To achieve a balanced distribution of topic ‘difficulty’, the topics were passed through a baseline retrieval system and topics with differing numbers of relevant images (as found by the baseline system), as well as topics with differing results when the baseline systems employed textual or visual evidence, were selected. Difficult topics usually convey complex semantics (e.g., “Chernobyl disaster ruins”); whereas easier topics have a clearly defined conceptual focus (e.g., “blue flower”). Image examples were provided for each topic to support the investigation of multimodal approaches. To further encourage multimodal approaches, the number of example images was significantly increased in 2011 (to 4.84 versus 1.68, 1.7, and 0.61 in previous years), which allowed participants to build more complex visual query models. The collections also incorporated additional visual

<sup>8</sup>[www.exalead.com/search](http://www.exalead.com/search).

**Table 1** Wikipedia image retrieval collections 2008–2011

|                                | 2008                 | 2009          | 2010                                   | 2011               |
|--------------------------------|----------------------|---------------|--|--------------------|
| Number of images in collection | 151,519              |               | 237,434                                |                    |
| Textual annotations            | description; caption |               | description; caption; comment; article |                    |
| Language(s)                    | English              |               | English, French, German                |                    |
| Topic development              | Collaborative        | Collaborative | Belga query logs                       | Exalead query logs |
| Number of topics               | 75                   | 45            | 70                                     | 50                 |
| Number of words/topic          | 2.64                 | 2.7           | 2.7                                    | 3.1                |
| Number of images/topic         | 0.61                 | 1.7           | 1.68                                   | 4.84               |

**Table 2** Participation, pooled runs, and relevance assessments for the Wikipedia image retrieval collections 2008–2011

|                                  | 2008         | 2009         | 2010          | 2011          |
|----------------------------------|--------------|--------------|---------------|---------------|
| Participants                     | 12           | 8            | 13            | 11            |
| Runs (textual/visual/multimodal) | 74 (36/5/33) | 57 (26/2/29) | 127 (48/7/72) | 110 (51/2/57) |
| Pool depth                       | 100          | 50           | 100           | 100           |
| Pool size/topic: average         | 1290         | 545          | 2659          | 1467          |
| Pool size/topic: minimum–maximum | 753–1850     | 299–802      | 1421–3850     | 764–2327      |
| Number of relevant images/topic  | 74.6         | 36.0         | 252.25        | 68.8          |

resources, such as extracted visual features, to encourage participation from groups that specialise in text retrieval. Table 1 summarises the main characteristics of the Wikipedia image retrieval collections 2008–2011.

Participation in the task consisted of 12 groups in 2008, 8 in 2009, 13 in 2010, and 11 in 2011. These groups submitted, respectively, 74, 57, 127, and 110 runs which were included in the pools to be assessed, using a pool depth of 100 in 2008, 2010, and 2011, and a pool depth of 50 in 2009. As Table 2 indicates the average pool size per topic varied over the years, even for the same pool depth. It was larger in 2010 and 2011 compared to 2008 due to many more runs being submitted, thus contributing more unique images to the pools. Also because the later collection was substantially larger, it is possible that the runs retrieved more diverse images.

During the first 3 years, volunteer task participants and the organisers performed the relevance assessments. To ensure consistency, a single person assessed the pooled images for each topic. As the number of volunteer assessors dropped significantly over the years, a crowdsourcing approach was adopted in 2011. The assessments were carried out by Amazon Mechanical Turk<sup>9</sup> workers using the CrowdFlower<sup>10</sup> platforms. Each worker assignment involved the assessment of five images for a single topic. To validate quality, each assignment contained one already annotated “gold standard” image among the five images so as to estimate workers’ accuracy. If a worker’s accuracy dropped below a threshold, their assessments were

<sup>9</sup><http://www.mturk.com>.

<sup>10</sup><http://crowdfower.com>.

excluded. To further increase accuracy, each image was assessed by three workers with the final assessment obtained through majority voting. Although this approach risks obtaining inconsistent results for the same topic, an inspection of the results did not reveal such issues and the ground truth creation was completed within a few hours, a marked difference to previous years.

The effectiveness of the submitted runs was evaluated using well-established evaluation measures, including Mean Average Precision (MAP), precision at fixed rank position ( $P@n$ ,  $n=10$  and  $20$ ), and R-precision (precision at rank position  $R$ , where  $R$  is the number of relevant documents). In addition, given that the relevance assessments are incomplete due to pooling, the binary preference (BPref) evaluation measure (Buckley and Voorhees 2004) was also adopted since it also accounts for unjudged documents. MAP was selected as the main evaluation measure given its higher inherent stability, informativeness, and discriminative power (Buckley and Voorhees 2000).

During the first 2 years of the Wikipedia image retrieval task (Tsirikika and Kludas 2009, 2010), the best performing approaches were text-based, with the multimodal approaches though continuously showing signs of improvement. Over the next 2 years (Popescu et al. 2010; Tsirikika et al. 2011b), the multimodal runs outperformed the monomodal approaches both in the overall ranking and also for the majority of teams that submitted both types of runs. A further analysis of the results also showed that most topics (particularly for 2011) were significantly better addressed with multimodal approaches, given also the increased number of query images. Moreover, the further fusion of results from multiple languages further improved the performance indicating the effectiveness of multilingual runs over monolingual ones due to the distribution of the information over the different languages.

Finally, a meta-analysis on the reliability and reusability of the test collections built during the 4 years the Wikipedia image retrieval task (Tsirikika et al. 2012) showed that 40–50 topics achieved stable rankings independent of the topic set, where a difference of at least 0.05 in MAP or BPref indicates that one run is consistently better than another. Our analysis also showed that a pool depth of 50 suffices to produce stable rankings for the given topic set sizes and that the created test collections can fairly rank most runs that do not contribute to the pool even though a few single runs might be heavily misjudged.

## 6 Discussions and Lessons Learned

Over the years ImageCLEF has contributed to advances in visual information retrieval through the provision of evaluation resources and organising large-scale evaluation activities. Various tasks have been offered to address different aspects of ad hoc multi-lingual retrieval of visual media. A detailed discussion of the evaluation activities and results can be found in Müller et al. (2010). Across the years results highlighted several recurring points: that combining information from visual



and textual features can improve retrieval; that bilingual retrieval that compares favourably with monolingual is possible even with limited text available (although dependent on annotation quality); query expansion and relevance feedback (using both textual and visual features) can boost retrieval performance, especially in cases with limited textual metadata; and the use of semantic knowledge bases can improve retrieval (e.g. for document expansion), especially in cases when text associated with the images is rich in entities.

Experiences with organising the ImageCLEF tasks and the challenges ensued were first summarised in Müller et al. (2007). These included: obtaining sufficient funding to organise activities; access to data to enable the distribution of re-usable evaluation resources; gaining sufficient interest and participation in events; gaining support from non-academic organisations; creating realistic tasks and user models to base testing on; and creating high quality ground truth data (given in particular the time and financial constraints). Some of the challenges and ways of addressing them are discussed below.

Building test collections for image retrieval tasks, similar to the ones described in this chapter, and with the goal to distribute them to the research community should take into account copyright issues of the images involved; the same applies also to any additional resources, such as any example images added to the topics. Furthermore, all users of image collections used in evaluation campaigns must be fully aware of original image license terms and adhere to them. This must be managed by the task organisers/collection distributors, e.g., identifying rights, informing participants and ensuring license terms are followed. In addition, we have found that prior to distributing image collections, it is also important to perform multiple checks on the integrity of images to support their use by participants.

With respect to developing suitable tasks, creating topics based on analysing query logs is more likely to generate more realistic topics that are closer to those most interesting to a general user population and can better simulate the needs of real-world users. Nevertheless, tasks that are distributed within evaluations, such as ImageCLEF, should contain topics of varying difficulty that also have a varying number of relevant documents in the collection; a baseline retrieval system that gives an overview of the collection content can help test the latter. Moreover, when the assessment uses pooling, the number of relevant documents should not be too large (e.g.,  $\leq 100$  images) to minimise the number of unjudged relevant documents that influence the stability of ranking. Our analysis of the datasets also indicates that ground truths should be created by pooling a large number of runs based on a variety of heterogeneous approaches that have the potential to retrieve diverse images, so as to be able to fairly rank new approaches.

Finally, in the case of the Wikipedia task the adoption of a crowdsourcing approach for performing relevance assessments was a positive experience, as assessments were carried out more quickly and accurately when more than one assessor was assigned to each topic. Crowdsourcing is nowadays employed widely for image annotation and retrieval tasks, most notably in the ImageNet initiative (Russakovsky et al. 2015) where high quality annotated datasets have been produced in a cost effective manner.

## 7 Conclusions

This chapter has discussed the activities in the context of the CLEF evaluation campaign regarding the multi-lingual retrieval of images from various sources in the non-medical domain. Four image retrieval tasks that ran intermittently as part of ImageCLEF over 9 years (2003–2011) were presented and the test collections built in the context of their activities were described. The overall goal of all the presented tasks has been to promote progress in large scale, multi-modal image retrieval via the provision of appropriate test collections that can be used to reliably benchmark the performance of different retrieval approaches using a metrics-based evaluation. The contributions and overall impact of these image retrieval tasks have been widely recognised in previous analyses (Tsitrika et al. 2011a, 2013) and this chapter further discusses some best practices based on the lessons learned when addressing the multiple challenges arising when building reliable and reusable test collections for image retrieval.

**Acknowledgements** We thank all those involved in helping to coordinate and organise ImageCLEF over the years and all those who have engaged with the track, providing submissions and presenting at CLEF.

## References

- Agrawal R, Gollapudi S, Halverson A, Jeong S (2009) Diversifying search results. In: Proceedings of the second ACM international conference on web search and data mining, WSDM'09. ACM, New York, pp 5–14. <http://doi.acm.org/10.1145/1498759.1498766>
- Arni T, Clough P, Sanderson M, Grubinger M (2009) Overview of the ImageCLEFphoto 2008 photographic retrieval task. In: Peters C, Deselaers T, Ferro N, Gonzalo J, Jones GJF, Kurimo M, Mandl T, Peñas A (eds) Evaluating systems for multilingual and multimodal information access: ninth workshop of the cross-language evaluation forum (CLEF 2008). Revised Selected Papers, Lecture notes in computer science (LNCS), vol 5706. Springer, Heidelberg, pp 500–511
- Boididou C, Papadopoulos S, Kompatsiaris Y, Schifferes S, Newman N (2014) Challenges of computational verification in social multimedia. In: Proceedings of the 23rd international conference on world wide web, WWW'14 companion. ACM, New York, pp 743–748. <http://doi.acm.org/10.1145/2567948.2579323>
- Buckley C, Voorhees EM (2000) Evaluating evaluation measure stability. In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR'00. ACM, New York, pp 33–40. <http://doi.acm.org/10.1145/345508.345543>
- Buckley C, Voorhees EM (2004) Retrieval evaluation with incomplete information. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR'04. ACM, New York, pp 25–32. <http://doi.acm.org/10.1145/1008992.1009000>
- Clarke CL, Kolla M, Cormack GV, Vechtomova O, Ashkan A, Büttcher S, MacKinnon I (2008) Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR'08. ACM, New York, pp 659–666. <http://doi.acm.org/10.1145/1390334.1390446>

- Clough P, Sanderson M (2004) The CLEF 2003 cross language image retrieval track. In: Peters C, Braschler M, Gonzalo J, Kluck M (eds) Comparative evaluation of multilingual information access systems: fourth workshop of the cross-language evaluation forum (CLEF 2003), revised selected papers. Lecture notes in computer science (LNCS), vol 3237. Springer, Heidelberg, pp 581–593
- Clough P, Müller H, Sanderson M (2005) The CLEF 2004 cross-language image retrieval track. In: Peters C, Clough P, Gonzalo J, Jones GJF, Kluck M, Magnini B (eds) Multilingual information access for text, speech and images: fifth workshop of the cross-language evaluation forum (CLEF 2004). Revised selected papers. Lecture notes in computer science (LNCS), vol 3491. Springer, Heidelberg, pp 597–613
- Clough P, Grubinger M, Deselaers T, Hanbury A, Müller H (2006a) Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In: Nardi A, Peters C, Vicedo JL, Ferro N (eds) CLEF 2006 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1172/>
- Clough PD, Sanderson M, Reid N (2006b) The eurovision St Andrews collection of photographs. SIGIR Forum 40(1):21–30. <http://doi.acm.org/10.1145/1147197.1147199>
- Clough P, Grubinger M, Deselaers T, Hanbury A, Müller H (2007) Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In: Peters C, Clough P, Gey FC, Karlgren J, Magnini B, Oard DW, de Rijke M, Stempfhuber M (eds) Evaluation of multilingual and multi-modal information retrieval: seventh workshop of the cross-language evaluation forum (CLEF 2006). Revised selected papers, Lecture notes in computer science (LNCS), vol 4730. Springer, Heidelberg, pp 223–256
- Del Bimbo A (1999) Visual information retrieval. Morgan Kaufmann, San Francisco
- Grubinger M, Clough P, Leung C (2006) The IAPR TC-12 benchmark for visual information search. IAPR Newsl 28(2):10–12
- Grubinger M, Clough P, Hanbury A, Müller H (2008) Overview of the ImageCLEFphoto 2007 photographic retrieval task. In: Peters C, Jijkoun V, Mandl T, Müller H, Oard DW, Peñas A, Petras V, Santos D (eds) Advances in multilingual and multimodal information retrieval: eighth workshop of the cross-language evaluation forum (CLEF 2007). Revised selected papers, Lecture notes in computer science (LNCS), vol 5152. Springer, Heidelberg, pp 433–444
- Husain SS, Bober M (2017) Improving large-scale image retrieval through robust aggregation of local descriptors. IEEE Trans Pattern Anal Mach Intell 39(9):1783–1796. <https://doi.org/10.1109/TPAMI.2016.2613873>
- Müller H, Deselaers T, Grubinger M, Clough P, Hanbury A, Hersh W (2007) Problems with running a successful multimedia retrieval benchmark. In: Proceedings of the third MUS-CLE/ImageCLEF workshop on image and video retrieval evaluation
- Müller H, Kalpathy-Cramer J, Eggel I, Bedrick S, Radhouani S, Bakke B, Khan CE, Jr, Hersh WR (2010) Overview of the CLEF 2009 medical image retrieval track. In: Peters C, Tsirikla T, Müller H, Kalpathy-Cramer J, Jones GJF, Gonzalo J, Caputo B (eds) Multilingual information access evaluation vol. ii multimedia experiments – tenth workshop of the cross-language evaluation forum (CLEF 2009). Revised selected papers, Lecture notes in computer science (LNCS). Springer, Heidelberg, pp 72–84
- Paramita ML, Grubinger M (2010) Photographic image retrieval. In: Müller H, Clough PD, Deselaers T, Caputo B (eds) ImageCLEF, experimental evaluation in visual information retrieval. Springer, Berlin, pp 141–162. [https://doi.org/10.1007/978-3-642-15181-1\\_8](https://doi.org/10.1007/978-3-642-15181-1_8)
- Paramita ML, Sanderson M, Clough P (2009a) Developing a test collection to support diversity analysis. In: Proceedings of redundancy, diversity, and IDR workshop-SIGIR, vol 9, pp 39–45
- Paramita ML, Sanderson M, Clough P (2009b) Diversity in photo retrieval: overview of the imageclefphoto task 2009. In: Workshop of the cross-language evaluation forum for European languages. Springer, Berlin pp 45–59
- Popescu A, Tsirikla T, Kludas J (2010) Overview of the wikipedia retrieval task at ImageCLEF 2010. In: Braschler M, Harman DK, Pianta E, Ferro N (eds) CLEF 2010 working notes, CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1176/>

- Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) Imagenet large scale visual recognition challenge. *Int J Comput Vis* 115(3):211–252. <https://doi.org/10.1007/s11263-015-0816-y>
- Santos RLT, Macdonald C, Ounis I (2015) Search result diversification. *Found Trends Inf Retr* 9(1):1–90. <http://dx.doi.org/10.1561/15000000040>
- Tsikrika T, Kludas J (2009) Overview of the WikipediaMM task at ImageCLEF 2008. In: Peters C, Deselaers T, Ferro N, Gonzalo J, Jones GJF, Kurimo M, Mandl T, Peñas A (eds) *Evaluating systems for multilingual and multimodal information access: ninth workshop of the cross-language evaluation forum (CLEF 2008)*. Revised selected papers, *Lecture notes in computer science (LNCS)*, vol 5706. Springer, Heidelberg, pp 539–550
- Tsikrika T, Kludas J (2010) Overview of the WikipediaMM task at ImageCLEF 2009. In: Peters C, Tsirikika T, Müller H, Kalpathy-Cramer J, Jones GJF, Gonzalo J, Caputo B (eds) *Multilingual information access evaluation vol. ii multimedia experiments – tenth workshop of the cross-language evaluation forum (CLEF 2009)*. Revised selected papers, *Lecture notes in computer science (LNCS)*. Springer, Heidelberg, pp 60–71
- Tsikrika T, Garcia Seco de Herrera A, Müller H (2011a) Assessing the scholarly impact of ImageCLEF. In: Forner P, Gonzalo J, Kekäläinen J, Lalmas M, de Rijke M (eds) *Multilingual and multimodal information access evaluation. Proceedings of the second international conference of the cross-language evaluation forum (CLEF 2011)*. *Lecture notes in computer science (LNCS)*, vol 6941. Springer, Heidelberg, pp 95–106
- Tsikrika T, Popescu A, Kludas J (2011b) Overview of the Wikipedia image retrieval task at ImageCLEF 2011. In: Petras V, Forner P, Clough P, Ferro N (eds) *CLEF 2011 working notes*. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1177/>
- Tsikrika T, Kludas J, Popescu A (2012) Building reliable and reusable test collections for image retrieval: the wikipedia task at imageclef. *IEEE MultiMedia* 19(3):24–33
- Tsikrika T, Larsen B, Müller H, Endrullis S, Rahm E (2013) The scholarly impact of CLEF (2000–2009). In: Forner P, Müller H, Paredes R, Rosso P, Stein B (eds) *Information access evaluation meets multilinguality, multimodality, and visualization. Proceedings of the fourth international conference of the CLEF initiative (CLEF 2013)*. *Lecture notes in computer science (LNCS)*, vol 8138. Springer, Heidelberg, pp 1–12
- Wan J, Wang D, Hoi SCH, Wu P, Zhu J, Zhang Y, Li J (2014) Deep learning for content-based image retrieval: a comprehensive study. In: *Proceedings of the 22Nd ACM international conference on multimedia, MM’14*. ACM, New York, pp 157–166. <http://doi.acm.org/10.1145/2647868.2654948>
- Westerveld T, van Zwol R (2007) Multimedia retrieval at inex 2006. *ACM SIGIR Forum* 41(1):58–63
- Zhai CX, Cohen WW, Lafferty J (2003) Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In: *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR’03*. ACM, New York, pp 10–17. <http://doi.acm.org/10.1145/860435.860440>

# Experiences from the ImageCLEF Medical Retrieval and Annotation Tasks



Henning Müller, Jayashree Kalpathy-Cramer, and Alba García  
Seco de Herrera

**Abstract** The medical tasks in ImageCLEF have been run every year from 2004–2018 and many different tasks and data sets have been used over these years. The created resources are being used by many researchers well beyond the actual evaluation campaigns and are allowing to compare the performance of many techniques on the same grounds and in a reproducible way. Many of the larger data sets are from the medical literature, as such images are easier to obtain and to share than clinical data, which was used in a few smaller ImageCLEF challenges that are specifically marked with the disease type and anatomic region. This chapter describes the main results of the various tasks over the years, including data, participants, types of tasks evaluated and also the lessons learned in organizing such tasks for the scientific community.

## 1 Introduction

ImageCLEF<sup>1</sup> started as the Cross-Language Image Retrieval Task in CLEF (Cross-Language Evaluation Forum<sup>2</sup>) in 2003 (Clough and Sanderson 2004; Clough et al.

---

<sup>1</sup><http://www.imageclef.org/>.

<sup>2</sup><http://www.clef-campaign.org/>.

---

H. Müller (✉)  
HES–SO Valais, Sierre, Switzerland  
e-mail: [henning.mueller@hevs.ch](mailto:henning.mueller@hevs.ch)

J. Kalpathy-Cramer  
MGH Martinos Center for Biomedical Imaging, Charlestown, MA, USA  
e-mail: [Kalpathy@nmr.mgh.harvard.edu](mailto:Kalpathy@nmr.mgh.harvard.edu)

A. García Seco de Herrera  
University of Essex, Colchester, UK  
e-mail: [alba.garcia@essex.ac.uk](mailto:alba.garcia@essex.ac.uk)

2010). A medical task was added in 2004 (Clough et al. 2005) and has been held every year since then (Kalpathy-Cramer et al. 2015). Several articles and books describe the overall evolution of the tasks and the various approaches that were used to create the resources and compare the results in much detail (Kalpathy-Cramer et al. 2015; Müller et al. 2010a). Similar to other campaigns such as TREC (Text Retrieval Conference) (Rowe et al. 2010) or TRECvid (The video retrieval task of the Text Retrieval Conference) (Thornley et al. 2011), an important scholarly impact was shown for both ImageCLEF (Tsikrika et al. 2011) and also the overall CLEF campaign (Tsikrika et al. 2013; Angelini et al. 2014). As the impact increases almost exponentially over the years it can be expected that the impact has grown even stronger since these studies were published in 2011 and 2013, respectively. Particularly the resources on medical data have been used by a large number of researchers, as many technical research groups find it hard to access medical data sets if they do not have a close collaboration with medical partners. As Open Science is generally supported strongly by funding organizations and universities, there is a whole field building around making data, tasks and code available and sharing these resources with other researchers. Such Open Science can strongly increase the impact of research projects as well, when sharing data and software.

The data sets and tasks in ImageCLEF have evolved over the years with data sets becoming generally larger and tasks more challenging and complex. Some clinically relevant data sets remain relatively small but this is simply linked to data availability and confidentiality, and also to the cost of annotation. An overall goal of ImageCLEF has always been to create resources that allow for multimodal data access, so combining visual and textual information and possibly structured data. Another objective was to develop tasks that are based on solid grounds and allow for an evaluation in a realistic scenario (Müller et al. 2007a). Log files of search systems have been used as well as example cases from teaching files (Müller et al. 2008b) to develop topics for retrieval system evaluation.

Scientific challenges were rare in the multimedia analysis or medical imaging field in the 1990s and 2000s compared to the information retrieval community, where they already started in the 1960s (Cleverdon et al. 1966; Jones and van Rijsbergen 1975). In medical imaging, systematic benchmarking really started with a few conferences adding challenges in the late 2000s (Heimann et al. 2009) and slightly earlier with the ImageCLEF benchmark but only for visual medical information retrieval. Since around 2010, most major conferences in the field of image analysis and machine learning propose scientific challenges similar to workshops that have been part of conference programs for many years and that usually take one or two days at these conferences. These conference challenges have strongly influenced the field, as many examples show (Menze et al. 2015; Jimenez-del-Toro et al. 2015). Many large data sets and also software are now being shared (via platforms such as GitHub) and used by a large number of researchers to compare techniques on the same grounds.

More recent changes are linked to research infrastructures where an objective was to move the algorithms towards the data rather than the data to the algorithms (Hanbury et al. 2012). This has many advantages when dealing with very large data sets,

confidential data, or sources that change and evolve quickly, when creating a fixed data collection is not practical. Several approaches have been presented for creating evaluation frameworks that allow the submission of source code, virtual machines or Docker containers (Jimenez-del-Toro et al. 2016; Gollub et al. 2012). More generally, such approaches are grouped under the term Evaluation-as-a-Service (EaaS<sup>3</sup>) (Hanbury et al. 2015), and are really an integrated way to share data, source code and computational infrastructures for research. A previous chapter in this volume discusses EaaS in more details.

This chapter analyzes the work done in the ImageCLEF medical tasks from 2004 until 2018 showing how tasks and techniques have evolved. It also gives many links to further resources, as an extremely detailed analysis of the participating techniques is not possible in such a short book chapter. The many references give good starting points for a more detailed analysis. The data sets created in ImageCLEF are also usually used for many years beyond the ImageCLEF challenges and these articles need to be analysed to show the real advances in system performance over the years.

This chapter is organized as follows: Sect. 2 describes the ImageCLEF tasks, the data sets and the participation. An overview of the main techniques that achieved best results is given in the last part of the section. The main lessons learned are described in Sect. 3 and conclusions are given in Sect. 4.

## 2 Tasks, Data and Participation in the ImageCLEF Medical Tasks over the Years

This section describes the evolution of the tasks over the years, starting with the types of tasks proposed, the data types used, data size available and the participation in the task. A short discussion of the main techniques leading to best results is given.

### 2.1 Overview of the Medical Tasks Proposed

This section analyzes past data and resources created in the medical tasks of ImageCLEF that have been organized for 15 years. The analysis is based on the overview articles of these years (Clough et al. 2005, 2006; Müller et al. 2008a, 2009, 2010b, 2012, 2006; Radhouani et al. 2009; Kalpathy-Cramer et al. 2011; García Seco de Herrera et al. 2013, 2015, 2016a; Dicente Cid et al. 2017b; Eickhoff et al. 2017) and is summarized in Table 1.

It can be seen that the first years of ImageCLEF offered mainly general retrieval and then classification tasks. In 2010, a case-based retrieval task that is closer to clinical applications was proposed. In 2014, a first task related to a clinically-

---

<sup>3</sup><http://www.eaas.cc/>.

**Table 1** Overview of the various tasks that have been performed over the years, ranging from general tasks in the beginning to some disease-oriented task later on that are marked as such

| Task type                     | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 |
|-------------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| Image-based retrieval         | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    |      |      |      |      |      |
| Image type classification     |      | x    | x    | x    | x    | x    |      |      |      |      |      |      |      |      |      |
| Case-based retrieval          |      |      |      |      |      |      | x    | x    | x    | x    |      |      |      |      |      |
| Image modality classification |      |      |      |      |      |      | x    | x    | x    | x    |      |      |      |      |      |
| Subfigure classification      |      |      |      |      |      |      |      |      |      |      |      | x    | x    |      |      |
| Compound figure detection     |      |      |      |      |      |      |      |      |      |      |      | x    | x    |      |      |
| Multi-label classification    |      |      |      |      |      |      |      |      |      |      |      | x    | x    |      |      |
| Compound figure separation    |      |      |      |      |      |      |      |      |      | x    |      | x    | x    |      |      |
| Liver CT annotation           |      |      |      |      |      |      |      |      |      |      | x    | x    |      |      |      |
| Caption prediction            |      |      |      |      |      |      |      |      |      |      |      | x    | x    | x    | x    |
| Tuberculosis classification   |      |      |      |      |      |      |      |      |      |      |      |      |      | x    | x    |
| Visual question answering     |      |      |      |      |      |      |      |      |      |      |      |      |      |      | x    |

relevant set of diseases was introduced (annotation of liver CT images with semantic categories of lesions) and since 2017 a tuberculosis task is similarly related to a real clinical application and need (looking at tuberculosis type and drug resistances of the bacteria in the images alone). Many of the later tasks were much more complex and required not only information retrieval competencies and features extraction from images but really targeted approaches towards extracting knowledge from the images. Research groups without a close link with health specialists often reported that it was challenging to estimate performance of their tools. A user analysis of retrieval based on images in the medical open access literature showed that research tasks are required that enrich meta data on images in the literature, as basically no information describing the images is available. The type of image (for example x-ray, CT, MRI, light microscopy image) can be used to filter images before visual image similarity retrieval is employed, as it can strongly focus the search and also use image type-dependent visual features. Such meta data in the images and also filtering are required to build retrieval applications based on the cleaned data. Compound figures are another challenge in the biomedical literature as many journal figures contain several subfigures with varying content and relationships among them because some journals limit the number of figures and this pushes authors to add more content into few figures. Such figures can have subfigures of different types and thus also have parts with the visual appearance of several sub-categories. With the exponential growth of the biomedical literature this can also be considered a priority area for the future, as images are available in almost unlimited quantities (growing exponentially) and getting ground truth is a main challenge. Crowdsourcing has been used for this (Foncubierta-Rodríguez and Müller 2012) (see Sect. 2.3).

An example topic with a query in three languages and image examples for the retrieval task in 2005 is shown in Fig. 1.





**Fig. 1** A query requiring more than visual retrieval but visual features can provide hints to good results (taken from ImageCLEF 2005)

## 2.2 Data Sets and Constraints for Medical Data

One of the major challenges in medical data analysis is the availability of large-scale resources. Any medical data usage in health institutions needs to be confirmed by local ethics committees and usually requires a targeted application with a clinical application that cannot be modified without changing the ethics agreement. This often limits the size and availability of medical data and ethics committees may completely restrict sharing data, so analyses can only be executed locally on the data. Exceptions are medical teaching files that are created with ethics approval and also the biomedical literature that contains many images that were acquired with ethics approval and are then made available publicly. These two facts also drove the data sets in the medical ImageCLEF tasks. Table 2 shows an overview of the types of images and the number of images or cases that are available in each of the years of ImageCLEF.

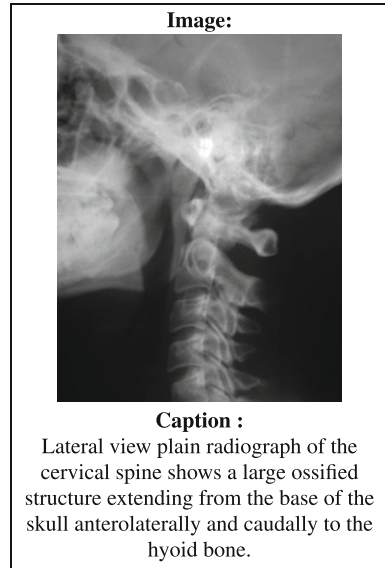
Whereas for most tasks the data set size is the number of images, for the tuberculosis task this is the number of volumes. Each volume then consists of around 150–200 slices or images. This explains the seemingly small size, even though the complexity of the tasks has significantly grown with the 3D data set need to be analyzed.

Most data sets are from the biomedical literature because this can make sharing data sets easier. Whereas the initial database of images from radiology journals was

**Table 2** Overview of the data sets that were created over the years for the various tasks

| Year | Task type                  | Resource                                | No images |
|------|----------------------------|---|-----------|
| 2004 | Image retrieval            | Teaching files, CasImage                | 8725      |
| 2005 | Image retrieval            | CasImage, PEIR, MIR, PathoPic           | 50,000    |
|      | Annotation                 | Radiographies of IRMA                   | 9000      |
| 2006 | Retrieval                  | CasImage, PEIR, MIR, PathoPic           | 50,000    |
|      | Annotation                 | Radiographics of IRMA                   | 11,000    |
| 2007 | Retrieval                  | myPACS, CORI added                      | 66,636    |
|      | Annotation                 | Radiographies of IRMA                   | 12,000    |
| 2008 | Retrieval                  | RSNA                                    | 66,000    |
|      | Annotation                 | Radiographies of IRMA                   | 12,076    |
| 2009 | Retrieval                  | RSNA                                    | 74,902    |
|      | Annotation                 | Radiographies of IRMA                   | 12,677    |
| 2010 | Retrieval                  | RSNA                                    | 77,506    |
|      | Case retrieval             | RSNA                                    | 77,506    |
|      | Classification             | RSNA modality classification            | 5010      |
| 2011 | Image retrieval            | PMC subset 1                            | 231,000   |
|      | Case retrieval             | PMC subset 1                            | 231,000   |
|      | Classification             | PMC subset 1 modality class.            | 2000      |
| 2012 | Image retrieval            | PMC subset 2                            | 300,000   |
|      | Case retrieval             | PMC subset 2                            | 300,000   |
|      | Classification             | PMC subset 2 modality class.            | 2000      |
| 2013 | Image retrieval            | PMC subset 2                            | 300,000   |
|      | Case retrieval             | PMC subset 2                            | 300,000   |
|      | Classification             | PMC subset 2 modality class.            | 5483      |
|      | Compound figure separation | PMC                                     | 2967      |
| 2014 | Annotation                 | Liver CT annotation dataset             | 60        |
| 2015 | Compound figure detection  | PMC subset 3                            | 20,867    |
|      | Compound figure separation | PMC subset 3 figure separation          | 6784      |
|      | Multi-label                | PMC subset 3 multi-label classification | 1568      |
|      | Classification             | PMC subset 3 subfigure classification   | 6776      |
|      | Clustering                 | Medical clustering                      | 5000      |
|      | Annotation                 | Liver CT annotation dataset             | 60        |
| 2016 | Compound figure detection  | PMC subset 4                            | 24,456    |
|      | Compound figure separation | PMC subset 4 figure sep.                | 8397      |
|      | Multi-label                | PMC subset 4 multi-label classification | 2651      |
|      | Classification             | PMC subset 4 subfigure classification   | 10,942    |
|      | Caption prediction         | PMC subset caption prediction 1         | 20,000    |
| 2017 | Caption prediction         | PMC subset caption prediction 2         | 184,614   |
|      | Concept detection          | PMC subset caption prediction 2         | 184,614   |
|      | Classification             | Tuberculosis dataset—MDR                | 444       |
|      | Resistance detection       | Tuberculosis dataset—TBT                | 801       |
| 2018 | Caption prediction         | PMC subset caption prediction 3         | 232,305   |
|      | Concept detection          | PMC subset caption prediction 3         | 232,305   |
|      | Classification             | Tuberculosis dataset—MDR                | 1513      |
|      | Resistance detection       | Tuberculosis dataset—TBT                | 495       |
|      | Severity scoring           | Tuberculosis dataset—SVR                | 279       |
|      | Visual question answering  | PMC subset VQA                          | 2866      |

**Fig. 2** Example of an image and its caption from the PubMed central dataset

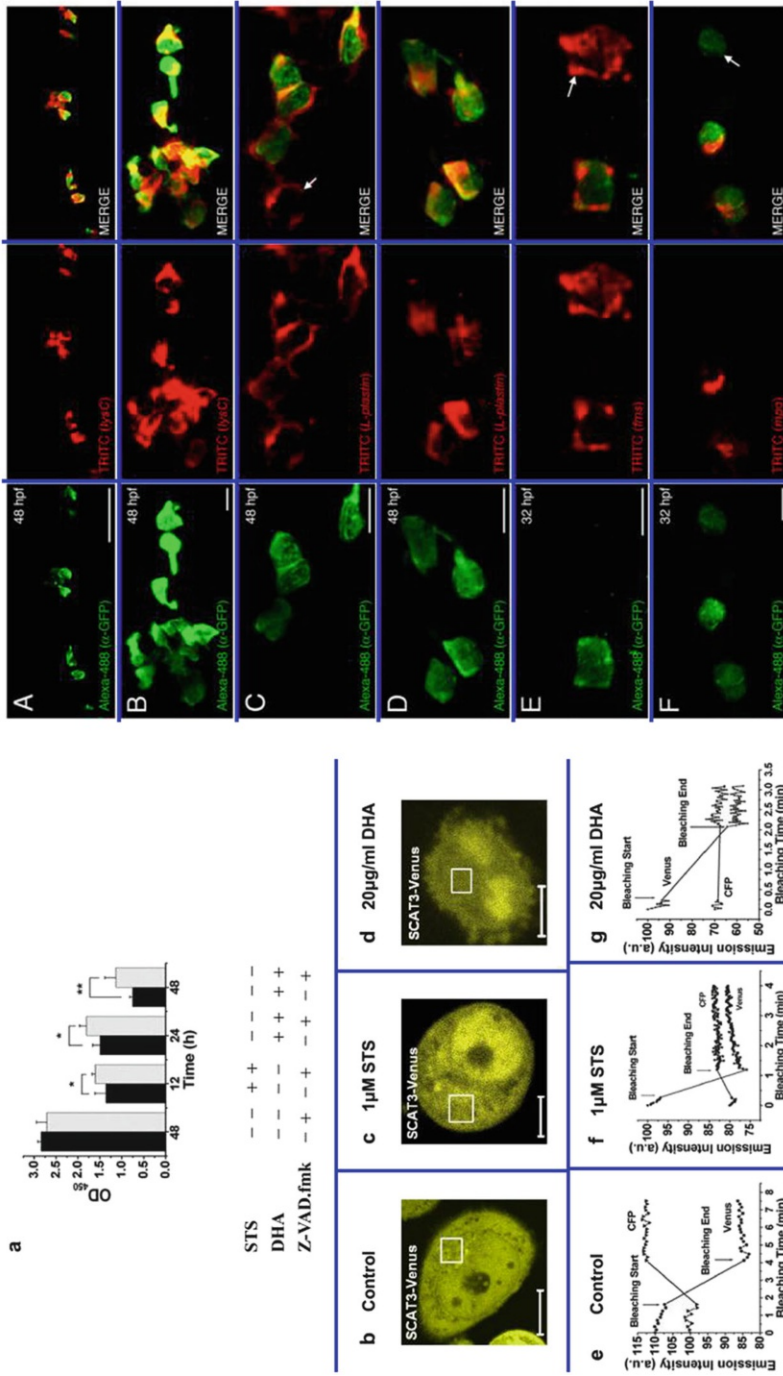


already filtered prior to using it and contained almost exclusively clinical images the images of PubMed Central (PMC) had a much larger variability. This variability can also be seen in Fig. 2 that shows an example of an image from the biomedical literature with its caption. Further examples are given later in the text.

One problem with images from the biomedical literature is shown in Fig. 3, which contains two compound figures and its parts that were automatically separated in this case. Compound figures are the majority of the content of PMC and their treatment thus has a massive impact on how the overall content of the biomedical literature can be exploited fully automatically. As subfigures can be of very different types the visual content is otherwise mixed and before attributing subfigures to a specific type they need to be separated.

### ***2.3 Relevance Judgements and Ground Truthing***

To develop a standard test bed for large and varied data sets, manually generated ground truth or relevance assessments (in the case of retrieval tasks) is basically always needed. Ground truth generation is costly, tedious and time-consuming. It is even more complex when specialists are required for tasks that can not be performed by the general public. Medical doctors are expensive and they often have no time for such ground truthing tasks. Sometimes, medical students can be used or other persons from health professions, and for simple and focused tasks crowdsourcing is a good option. For crowdsourcing, several relevance assessments are usually collected and used to eliminate incoherent results and to obtain a high



**Fig. 3** Examples of successful automatic separation results of compound figures (blue lines separate the subfigures)

quality (Clough and Sanderson 2013; García Seco de Herrera et al. 2014, 2016b). Several assessors agreeing usually means that the results are fine but there also need to be strategies to combine several judgements where disagreement exists.

For retrieval tasks a full judgement of an entire collection is not possible and thus a pooling technique is frequently used (Jones and van Rijsbergen 1975). Basically all image retrieval experiments in ImageCLEF on larger datasets use pooling, so the top  $N$  results of all participating runs are put together into a pool per topic and only these documents are judged for relevance. For classification usually the entire collection is classified manually and thus the data sets are often smaller than for the retrieval tasks. With sufficient training of very specific tasks also non-medical staff can be used for the classifications or relevance assessments, so crowdsourcing with quality control is also possible.

For ImageCLEF, the ground truth was in the first years generated by medical doctors, also because the collections were much smaller (500 images in 2004). Then, health science students could be hired, of which many were physicians. This was only possible thanks to funding that was available via related research projects. Limited funding was then used for crowdsourcing. In the past few years tasks based on data from the literature were created where no manual ground truthing was required (for example the caption prediction task) or data sets were obtained where the ground truth was already available (as in the tuberculosis task). Sometimes also a combination of approaches was used, partly with manual judgements and partly with crowdsourcing. More details can be found in the overview papers of the respective tasks.

## 2.4 Evaluation Measures

The relevance assessments or ground truth are used to quantify system effectiveness (Clough and Sanderson 2013). Many evaluation measures can be used to assess performance of retrieval or classification tasks based on the number of relevant documents (Buyya and Venugopal 2005). The `trec_eval`<sup>4</sup> package is used as a standard tool for text retrieval and it extracts all relevant measures of ImageCLEF and most other benchmarks. Usually early precision and MAP (Mean Average Precision) are used as lead measures. Sometimes BPref (Binary Preference) is added as a measure that takes into account documents that were not judged in the pooling process.

Accuracy is most commonly used to assess classification tasks. When the class distribution is very unbalanced there are also several other measures that are important, for example the geometric mean of the performance on all classes. This highlights a good performance on all classes and not a concentration on good results for a few majority classes, which would favor a good accuracy in this case. For

---

<sup>4</sup>[http://trec.nist.gov/trec\\_eval/](http://trec.nist.gov/trec_eval/).

medical tasks, specificity (true negative rate,  $1 - \text{false positive rate}$ ) and sensitivity (true positive rate or recall) are also very frequently used measures. These two measures allow to discriminate between whether it is important catch all patients with a condition or whether it is more important to limit false positives. Each community thus has its own measures and it is always important to show several measures to analyse different aspects of the performance of participating systems.

However, assessing tasks such as compound figure separation is challenging. In this case a new evaluation approach was developed. The evaluation required to have a minimum overlap for the subfigure division between the ground truth and the data supplied by the groups in their runs (García Seco de Herrera et al. 2013). This allowed for some margin in terms of the separating lines, which is important as there is not one single optimal solution and the judges doing the ground truthing had an important amount of subjectivity.

In general, it is important to have more than one performance measure and ranking to really evaluate several aspects of the participating techniques and to not concentrate all techniques into optimizing a single measure.

## 2.5 *Participants and Submissions*

In Table 3 the number of groups that registered for a task and the number of groups that finally submitted results are listed. For some of the years the exact registration numbers were not mentioned in the overview papers and thus we cannot reproduce them anymore. Thus, we used square brackets for these and used the number of submissions as a lower bound of the participation.

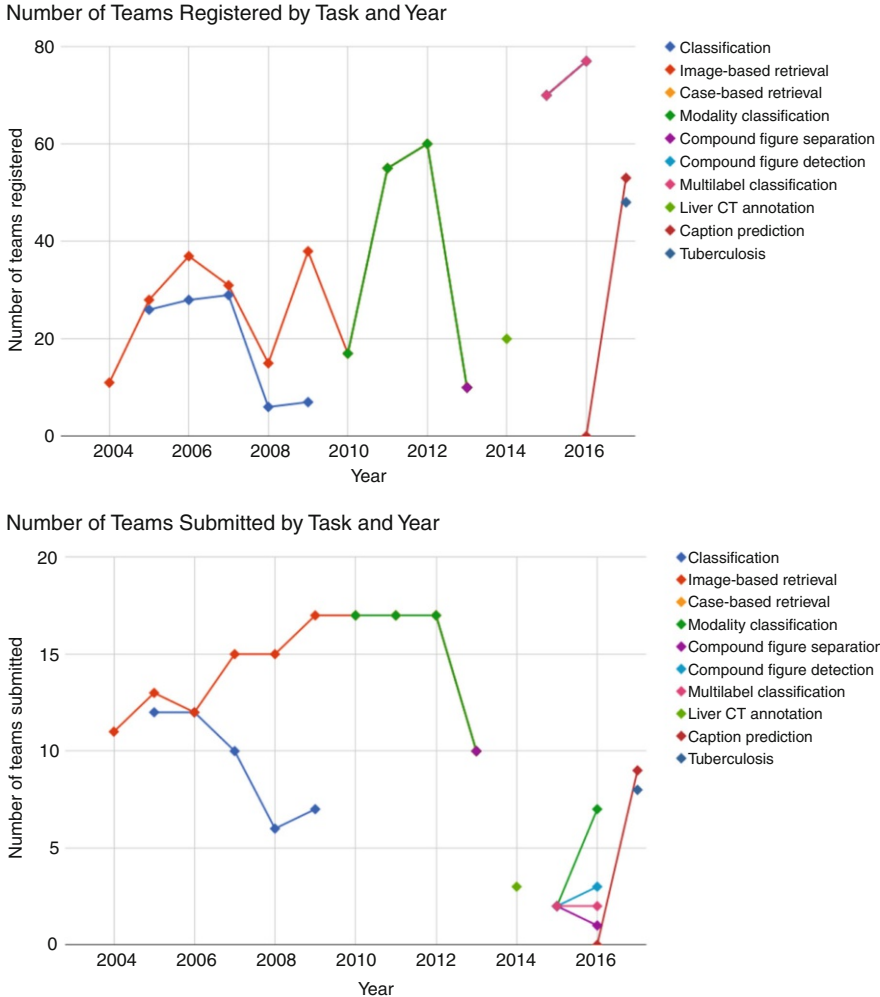
There has been a general increase in the participation over the years, but many new tasks take 1 or 2 years to obtain higher numbers because researchers need to adapt to specific tasks in their research projects. The number of submissions on the other hand has been lower in recent years where many new and more complex tasks were introduced that go beyond simple text retrieval or image classification.

Figure 4 shows the evolution of research groups that registered for the Image-CLEF medical tasks on a per task basis and in the second graphic those groups that submitted results. We can see that the long running tasks had a large number of actual submissions whereas the more recent tasks that have only been organized for 1–2 years have relatively few submissions. The number of registrations actually had some peaks in recent years and it seems to increase over the years in a relatively stable fashion. On the other hand, the percentage of registered users actually submitting results has decreased over this period. Possibly, this can be attributed to a larger availability of benchmarks and data sets for researchers to choose from.

**Table 3** Overview of the participation in the tasks over the years, “[ ]” denotes years when the exact numbers of registered users are not known (only the number of those submitting results) and “\*” highlights a task where in the end no group submitted results, which in combination with “[ ]” means that nothing concrete can be said about participation

| Year | Task                       | Registered | Submitted |
|------|----------------------------|------------|-----------|
| 2004 | Image-based retrieval      | [11]       | 11        |
| 2005 | Image-based retrieval      | 28         | 13        |
|      | Classification             | 26         | 12        |
| 2006 | Image-based retrieval      | 37         | 12        |
|      | Classification             | 28         | 12        |
| 2007 | Image-based retrieval      | 31         | 15        |
|      | Classification             | 29         | 10        |
| 2008 | Image-based retrieval      | [15]       | 15        |
|      | classification             | [6]        | 6         |
| 2009 | Image-based retrieval      | 38         | 17        |
|      | Classification             | [7]        | 7         |
| 2010 | Image-based retrieval      | [17]       | 17        |
|      | Case-based retrieval       | [17]       | 17        |
|      | Modality classification    | [17]       | 17        |
| 2011 | Image-based retrieval      | 55         | 17        |
|      | Case-based retrieval       | 55         | 17        |
|      | Modality classification    | 55         | 17        |
| 2012 | Image-based retrieval      | 60         | 17        |
|      | Case-based retrieval       | 60         | 17        |
|      | Modality classification    | 60         | 17        |
| 2013 | Image-based retrieval      | [10]       | 10        |
|      | Case-based retrieval       | [10]       | 10        |
|      | Modality classification    | [10]       | 10        |
|      | Compound figure separation | [10]       | 10        |
| 2014 | Liver CT annotation        | 20         | 3         |
| 2015 | Modality classification    | 70         | 2         |
|      | Compound figure separation | 70         | 2         |
|      | Compound figure detection  | 70         | 2         |
|      | Multi-label classification | 70         | 2         |
|      | Liver annotation           | 51         | 1         |
| 2016 | Modality classification    | 77         | 7         |
|      | Compound figure separation | 77         | 1         |
|      | Compound figure detection  | 77         | 3         |
|      | Multi-label classification | 77         | 2         |
|      | Caption prediction         | [0]*       | 0         |
| 2017 | Caption prediction         | 53         | 9         |
|      | Tuberculosis               | 48         | 8         |
| 2018 | Caption prediction         | 46         | 8         |
|      | Tuberculosis               | 33         | 11        |
|      | Visual question answering  | 48         | 5         |

The participants list can also include the task organizer if the team registered



**Fig. 4** Number of groups registered for ImageCLEF per task (figure at the top) and the number of groups that actually submitted runs (figure at the bottom) over all the years

## 2.6 Techniques Used

Whereas first techniques applied in ImageCLEF used mainly simple texture (Gabor filters, Tamura, co-occurrence matrices) and color (color histograms) features extracted from the images in combination with often simple distance measures such as k-nearest neighbors (k-NN), there were also first tests with combinations of text retrieval and visual retrieval techniques (Müller et al. 2005). In general, techniques can clearly be separated into text retrieval and visual analysis techniques, where text retrieval usually led to much better results for the retrieval tasks, whereas in



classification tasks often the visual results were better. Best results in the first years (2004–2007) were often obtained using simple feature modelling techniques similar to visual words (Deselaers et al. 2005) or Fisher vectors based on patches in the images and not the global image content alone. These techniques had very good results for several years until more elaborate machine learning approaches such as support vector machines (SVMs) really improved outcomes for all classification tasks (Tommasi et al. 2010). Details of all techniques are impossible to be described here. Often similar techniques led in some cases to very good results and in other cases to poor results depending on how well the techniques were really optimized.

Feature fusion remained another area where many approaches were tested (Depeursinge and Müller 2010). Often rank-based fusion led to better results than score-based fusion with text retrieval and image retrieval following very different distributions in terms of absolute similarity scores. Both early and late fusion sometimes led to best results, so this might really depend on the exact data and application scenario. Another major advance in terms of techniques was the use of Fisher vectors (Clinchant et al. 2010) that led to best results in several competitions.

In the past 3 years most successful techniques use deep learning approaches (Koitka and Friedrich 2016; Stefan et al. 2017) for most tasks. This holds true for almost all classification challenges but also more complex scenarios such as compound figure separation. Extraction of features from deep learning with classical classification techniques were also tested with success. There are several rather specific techniques that led to best results in focused tasks such as the tuberculosis task in 2017 (Dicente Cid et al. 2017a). Here, a graph model was used that obtained best results in prediction multiple drug resistances. This can be attributed to the modeling of known knowledge on lung anatomy and distribution of disease, which would require a very large number of cases to learn the model with deep learning. Using more handcrafted features can model this existing knowledge.

### 3 Lessons Learned and Mistakes to Avoid

In Müller et al. (2007b) an early summary already gave several important lessons learned from running the first 4 years of the medical tasks in ImageCLEF. Since then, many things have changed with scientific challenges really becoming a standard tool in medical imaging and computer vision. Particularly the diversity of the medical tasks in ImageCLEF has increased massively over the years.

The main success factor for any scientific challenge is really to *create a community* around the task and engage participants in the entire process. This creates a positive energy and attracts other participants and particularly motivates to pursue and submit results in the end. Strong participation by peers also increases the number of groups actually submitting results. This number is often small and in the range of 20–30% of the groups that initially registered. It ensures that a task is not only run a single year but several years in a row. Tsirikika et al. (2011) show that

for most of the tasks, there is a peak in terms of scholarly impact in their second or third year of operation, then followed by a slow stalling or even decline in impact if the tasks are not changed substantially. Running the same task for several years can lead to continuous improvement of the participating approaches. An important aspect is also to keep a continuous test set over the years to also measure absolute improvements of the techniques over time but this is often more difficult to realize.

Another important part that is linked to the community aspects is the general *communication* with participants. This is essential to keep participants or interested researchers updated on all details and the status of the competition at all times. The main entry point for all information in ImageCLEF is the web page that is regularly updated and contains all information on the tasks with details on data, task creation and performance measures, also of previous years. Results of the challenge are also published here. A registration system manages data access that requires the signature of an end user agreement. The registration system also allows to upload runs and all runs are automatically checked to be in the right format and only contain valid identifiers. This strongly reduces the work of organizers to check the submitted runs for mistakes, which was a common problem in the first years of ImageCLEF. A mailing list with all registered participants makes it possible to address all participants with targeted information, for example of deadline changes. As past participants can remain on the list this is also a prime means for announcing new tasks or task ideas that can be discussed with researchers. In recent years the communication strategy increasingly includes social media. ImageCLEF has a Facebook page<sup>5</sup> and a Twitter account<sup>6</sup> and these are also used to address participants. Part of this may be redundant but it makes it possible to reach all participants via a variety of channels. LinkedIn has also been used in recent years to advertise the tasks and broaden the participant base via focused groups in the area. In 2018, a new registration system based on the open source tool crowdAI<sup>7</sup> was implemented. This tool gives new possibilities, for example to not only have a final workshop where results are compared but a continuous leader board that is active also after the competition finishes and where groups can upload and compare their results in a continuous way. The use of EaaS approaches with code submission is also possible with such an infrastructure but currently not used by us.

Having a common *publication* that describes the data set, the creation of ground truth and that compares the results of all submitted results is another aspect that is important for reproducibility of the results and also for keeping the data accessible long term and having it used in a clear evaluation context. For this it is essential to have a description of the runs of the participants, so not only performance measures can be compared but also the techniques that lead to a specific performance. In the past it was often the case that best and worst approaches were using almost the same

---

<sup>5</sup><https://www.facebook.com/profile.php?id=106932209431744>.

<sup>6</sup><https://twitter.com/imageclef?lang=en>.

<sup>7</sup><http://www.crowdai.org/>.

techniques but that small modifications had important effects on the outcome and for this reason a formal description of all techniques is essential.

Linked to the publications and an analysis of the results is the organization of a common *workshop*. At the workshop, participants can present the most interesting results of each task and can then compare the approaches and outcomes to find better ways to improve results in the future. This can foster collaborations between participants even if in the past only a few collaborations between research groups have evolved from the discussions in the meeting. The workshop has open discussion sections each year to plan tasks and also evaluate procedures for the future and thus integrate feedback into improving the tasks. This is linked to a community feeling among participants and can clearly improve motivation if handled well. It is important to transparently discuss all details, so the rankings are based on solid grounds.

To tackle *current research challenges* is also important, as universities which are the main participants of the tasks all depend on funding and this is usually assigned based on calls for topics that are currently hot research topics. If topics really are novel then a PhD student can for example engage in several years of work on such challenges in a efficient way, where they can compare results to others and rely on the same setting and data. Usually, challenges get harder each year, so the full potential of the techniques can really be tested over time.

No collection or setup for an evaluation campaign is free of errors and thus it is essential to have structures and manpower to *fix errors* and mistakes in the data and the evaluations quickly, as soon as participants report them. This creates confidence in the evaluation campaign and makes sure that meaningful results can be obtained in the end. The capacity to fix errors and run a professional campaign is also linked with obtaining good *funding* for such challenges. Most often, only research funding is available and infrastructures that create data sources, maintaining basic services for benchmarks and a physical infrastructure are harder to fund, even though scientific impact in terms of citations can be higher for data papers than for technical papers, as many researchers base their work on this. Without funding, a certain professionalism can be lost as all organizers engage in their free time as volunteers. With respect to ground truthing, whether manual annotations of the data or relevance judgments, it is important to have funding, even when relatively cheap options such as crowdsourcing are used.

An objective of ImageCLEF has always been to be *complementary* with other evaluation tasks, in other conferences (for example TRECvid) or also inside the CLEF labs, such as LifeCLEF and CLEFeHealth. Such a complementarity ensures a clear positioning of the tasks and thus also a good participation. There have also been suggestions to organize ImageCLEF with existing conferences in computer vision or machine learning, as most tasks at CLEF have been focusing rather on text analysis and retrieval. We have had collaborations with other conferences in the past but feel that CLEF is a good forum for multimodal interdisciplinary research.

## 4 Discussion and Conclusions

The chapter gives an overview of 15 years of scientific challenges in the medical ImageCLEF tasks. It is clear that no extremely detailed analysis can be given on all lessons learned and results obtained for 15 years in only a few pages. This text mainly focuses on overviews of how the data, the tasks and the techniques evolved over the years. Then, we highlight the lessons learned and several success factors that were identified via discussions among the organizers and also with participants.

With Open Science now gaining momentum in almost all fields related to data science many challenges have been organized at conferences and workshops. Many of the challenges are similar in nature or in the data used. With an increasing number there can be fewer participants in every single challenge, which reduces the impact of every single challenge. Professional platforms such as Kaggle<sup>8</sup> have also changed the field of scientific challenges but leading researchers to commercial challenges, where prize money is available instead of publications at purely scientific challenges. The targets are in this case very different, so not so much on understanding the techniques but really on tuning existing techniques. There is clearly a large market for data science challenges and such complementary approaches will likely coexist in the future.

Whereas professional challenges with prize money often do not focus on documenting techniques of the runs submitted in detail and understanding the actual techniques they push towards optimal performance. Scientific challenges, sometimes also called coepetitions (in between a competition and a cooperation), on the other hand aim at reproducible science that documents all experiments that were run and also concentrates on the interestingness of approaches and algorithms and not only pure performance. We feel that this contributes to better understanding techniques and having a long term optimization of approaches. Cheating in such scientific challenges seems less likely than when prize money is involved, even though it still needs to be checked that results are compared in a fair way.

There are clearly many next steps that can be taken for scientific challenges. It is important to keep a workshop where participants meet but also keeping past challenges and data open for new submissions is important, so best results can be tracked and compared over a longer period of time. Fostering more collaboration is one of our important objectives that has not been easy to reach. Maybe components based, for example, on Docker containers can be used in automatic work flows and help to make component sharing easier among researchers. With machine learning going increasingly towards deep learning it also becomes possible to explore large data sets with various levels of annotations, so for example, high level manual annotations but also noisy automatic annotations that could augment the training data, for example with silver corpuses (Krenn et al. 2016).

---

<sup>8</sup><http://www.kaggle.com/>.

**Acknowledgements** We would like to thank the various funding organizations that have helped make ImageCLEF a reality (EU FP6 & FP7, SNF, RCSI, Google and others) and also all the volunteer researchers who helped organize the tasks. Another big thank you goes to the data providers that assured that medical data could be shared with the participants. A final thanks to all participants who work on the tasks and provide us with techniques to compare and with lively discussions at the ImageCLEF workshops.

## References

- Angelini M, Ferro N, Larsen B, Müller H, Santucci G, Silvello G, Tsirikla T (2014) Measuring and analyzing the scholarly impact of experimental evaluation initiatives. In: Italian research conference on digital libraries
- Buyya R, Venugopal S (2005) A gentle introduction to grid computing and technologies. *CSI Commun* 29(1):9–19
- Cleverdon C, Mills J, Keen M (1966) Factors determining the performance of indexing systems. Tech. Rep., ASLIB Cranfield Research Project, Cranfield
- Clinchant S, Csúrka G, Ah-Pine J, Jacquet G, Perronnin F, Sánchez J, Minoukadeh K (2010) Xrce’s participation in wikipedia retrieval, medical image modality classification and ad-hoc retrieval tasks of imageclef 2010. In: Working notes of the 2010 CLEF workshop
- Clough P, Sanderson M (2004) The CLEF 2003 cross language image retrieval task. In: Proceedings of the cross language evaluation forum (CLEF 2003)
- Clough P, Sanderson M (2013) Evaluating the performance of information retrieval systems using test collections. *Information Research* 18(2). <http://informationr.net/ir/18-2/paper582.html#.XSXc5S2B06g>
- Clough P, Müller H, Sanderson M (2005) The CLEF 2004 cross-language image retrieval track. In: Peters C, Clough P, Gonzalo J, Jones GJF, Kluck M, Magnini B (eds) Multilingual information access for text, speech and images: result of the fifth CLEF evaluation campaign. Lecture notes in computer science (LNCS), vol 3491, Springer, Bath, pp 597–613
- Clough P, Müller H, Deselaers T, Grubinger M, Lehmann TM, Jensen J, Hersh W (2006) The CLEF 2005 cross-language image retrieval track. In: Cross language evaluation forum (CLEF 2005). Lecture notes in computer science (LNCS). Springer, Berlin, pp 535–557
- Clough P, Müller H, Sanderson M (2010) Seven years of image retrieval evaluation. Springer, Berlin, pp 3–18
- Depeursinge A, Müller H (2010) Fusion techniques for combining textual and visual information retrieval. In: Müller H, Clough P, Deselaers T, Caputo B (eds) ImageCLEF, The Springer international series on information retrieval, vol 32. Springer, Berlin, pp 95–114
- Deselaers T, Weyand T, Keysers D, Macherey W, Ney H (2005) FIRE in ImageCLEF 2005: combining content-based image retrieval with textual information retrieval. In: Working notes of the CLEF workshop, Vienna
- Dicente Cid Y, Batmanghelich K, Müller H (2017a) Textured graph-model of the lungs for tuberculosis type classification and drug resistance prediction: participation in ImageCLEF 2017. In: CLEF2017 working notes. CEUR workshop proceedings, Dublin, CEUR-WS.org. <http://ceur-ws.org>
- Dicente Cid Y, Kalinovskiy A, Liauchuk V, Kovalev V, Müller H (2017b) Overview of ImageCLEFtuberculosis 2017 - predicting tuberculosis type and drug resistances. In: CLEF 2017 labs working notes. CEUR Workshop Proceedings. Dublin, CEUR-WS.org. <http://ceur-ws.org>
- Eickhoff C, Schwall I, García Seco de Herrera A, Müller H (2017) Overview of ImageCLEFcaption 2017 - the image caption prediction and concept extraction tasks to understand biomedical images. In: CLEF2017 working notes. CEUR workshop proceedings. Dublin, CEUR-WS.org. <http://ceur-ws.org>

- Foncubierta-Rodríguez A, Müller H (2012) Ground truth generation in medical imaging: a crowdsourcing based iterative approach. In: Workshop on crowdsourcing for multimedia
- García Seco de Herrera A, Kalpathy-Cramer J, Demner Fushman D, Antani S, Müller H (2013) Overview of the ImageCLEF 2013 medical tasks. In: Working notes of CLEF 2013 (cross language evaluation forum)
- García Seco de Herrera A, Foncubierta-Rodríguez A, Markonis D, Schaer R, Müller H (2014) Crowdsourcing for medical image classification. In: Annual congress SGMI 2014
- García Seco de Herrera A, Müller H, Bromuri S (2015) Overview of the ImageCLEF 2015 medical classification task. In: Working notes of CLEF 2015 (cross language evaluation forum)
- García Seco de Herrera A, Schaer R, Bromuri S, Müller H (2016a) Overview of the ImageCLEF 2016 medical task. In: Working notes of CLEF 2016 (cross language evaluation forum)
- García Seco de Herrera A, Schaer R, Antani S, Müller H (2016b) Using crowdsourcing for multi-label biomedical compound figure annotation. In: MICCAI workshop Labels. Lecture notes in computer science. Springer, Berlin
- Gollub T, Stein B, Burrows S, Hoppe D (2012) Tira: configuring, executing, and disseminating information retrieval experiments. In: 2012 23rd international workshop on database and expert systems applications (DEXA). IEEE, Piscataway, pp 151–155
- Hanbury A, Müller H, Langs G, Weber MA, Menze BH, Fernandez TS (2012) Bringing the algorithms to the data: cloud-based benchmarking for medical image analysis. In: CLEF conference. Lecture notes in computer science. Springer, Berlin
- Hanbury A, Müller H, Balog K, Brodt T, Cormack GV, Eggel I, Gollub T, Hopfgartner F, Kalpathy-Cramer J, Kando N, Krithara A, Lin J, Mercer S, Potthast M (2015) Evaluation-as-a-service: overview and outlook. ArXiv 1512.07454
- Heimann T, Van Ginneken B, Styner M, Arzhaeva Y, Aurich V, Bauer C, Beck A, Becker C, Beichel R, Bekes G, et al (2009) Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Trans Med Imag* 28(8):1251–1265
- Jimenez-del-Toro O, Hanbury A, Langs G, Foncubierta-Rodríguez A, Müller H (2015) Overview of the VISCERAL retrieval benchmark 2015. In: Multimodal retrieval in the medical domain: first international workshop, MRMD 2015, Vienna, Austria, March 29, 2015, Revised selected papers. Lecture notes in computer science, vol 9059. Springer, Berlin, pp 115–123
- Jimenez-del-Toro O, Müller H, Krenn M, Gruenberg K, Taha AA, Winterstein M, Eggel I, Foncubierta-Rodríguez A, Goksel O, Jakab A, Kontokotsios G, Langs G, Menze B, Salas Fernandez T, Schaer R, Wallejo A, Weber MA, Dicente Cid Y, Gass T, Heinrich M, Jia F, Kahl F, Kechichian R, Mai D, Spanier AB, Vincent G, Wang C, Wyeth D, Hanbury A (2016) Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL anatomy benchmarks. *IEEE Trans Med Imag* 35(11):2459–2475
- Jones KS, van Rijsbergen C (1975) Report on the need for and provision of an ideal information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge
- Kalpathy-Cramer J, Müller H, Bedrick S, Eggel I, García Seco de Herrera A, Tsirikla T (2011) The CLEF 2011 medical image retrieval and classification tasks. In: Working notes of CLEF 2011 (cross language evaluation forum)
- Kalpathy-Cramer J, García Seco de Herrera A, Demner-Fushman D, Antani S, Bedrick S, Müller H (2015) Evaluating performance of biomedical image retrieval systems: overview of the medical image retrieval task at ImageCLEF 2004–2014. *Comput Med Imag Graph* 39:55–61
- Koitka S, Friedrich CM (2016) Traditional feature engineering and deep learning approaches at medical classification task of ImageCLEF 2016. In: CLEF2016 working notes. CEUR workshop proceedings. CEUR-WS.org, Évora
- Krenn M, Dorfer M, Jimenez-del-Toro O, Müller H, Menze B, Weber MA, Hanbury A, Langs G (2016) Creating a large-scale silver corpus from multiple algorithmic segmentations. Springer, Berlin, pp 103–115

- Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R, Lanczi L, Gerstner E, Weber MA, Arbel T, Avants BB, Ayache N, Buendia P, Collins DL, Cordier N, Corso JJ, Criminisi A, Das T, Delingette H, Demiralp C, Durst CR, Dojat M, Doyle S, Festa J, Forbes F, Geremia E, Glocker B, Golland P, Guo X, Hamamci A, Iftekharuddin KM, Jena R, John NM, Konukoglu E, Lashkari D, Mariz JA, Meier R, Pereira S, Precup D, Price SJ, Raviv TR, Reza SMS, Ryan M, Sarikaya D, Schwartz L, Shin HC, Shotton J, Silva CA, Sousa N, Subbanna NK, Szekely G, Taylor TJ, Thomas OM, Tustison NJ, Unal G, Vasseur F, Wintermark M, Ye DH, Zhao L, Zhao B, Zikic D, Prastawa M, Reyes M, Van Leemput K (2015) The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imag* 34(10):1993–2024
- Müller H, Geissbuhler A, Ruch P (2005) ImageCLEF 2004: combining image and multi-lingual search for medical image retrieval. In: Peters C, Clough P, Gonzalo J, Jones GJF, Kluck M, Magnini B (eds) *Multilingual information access for text, speech and images: result of the fifth CLEF evaluation campaign*. Lecture notes in computer science (LNCS), vol 3491. Springer, Bath, pp 718–727
- Müller H, Deselaers T, Lehmann T, Clough P, Kim E, Hersh W (2006) Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. In: Nardi A, Peters C, Vicedo JL, Ferro N (eds) *CLEF 2006 working notes*. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1172/>
- Müller H, Boyer C, Gaudinat A, Hersh W, Geissbuhler A (2007a) Analyzing web log files of the health on the net HONmedia search engine to define typical image search tasks for image retrieval evaluation. In: *MedInfo 2007*. Studies in health technology and informatics, Brisbane, vol 12. IOS Press, Amsterdam, pp 1319–1323
- Müller H, Deselaers T, Grubinger M, Clough P, Hanbury A, Hersh W (2007b) Problems with running a successful multimedia retrieval benchmark. In: *MUSCLE/ImageCLEF workshop 2007*, Budapest, pp 9–18
- Müller H, Deselaers T, Kim E, Kalpathy-Cramer J, Deserno TM, Clough P, Hersh W (2008a) Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: *CLEF 2007 proceedings*. Lecture notes in computer science (LNCS), Springer, Budapest, vol 5152, pp 473–491
- Müller H, Kalpathy-Cramer J, Hersh W, Geissbuhler A (2008b) Using medline queries to generate image retrieval tasks for benchmarking. In: *Medical informatics Europe (MIE2008)*. IOS Press, Gothenburg, pp 523–528
- Müller H, Kalpathy-Cramer J, Kahn CE, Jr, Hatt W, Bedrick S, Hersh W (2009) Overview of the ImageCLEFmed 2008 medical image retrieval task. In: Peters C, Giampiccolo D, Ferro N, Petras V, Gonzalo J, Peñas A, Deselaers T, Mandl T, Jones G, Kurimo M (eds) *Evaluating systems for multilingual and multimodal information access – 9th workshop of the cross-language evaluation forum*, Aarhus, Denmark. Lecture Notes in Computer Science (LNCS), vol 5706, pp 500–510
- Müller H, Clough P, Deselaers T, Caputo B (eds) (2010a) *ImageCLEF – experimental evaluation in visual information retrieval*. The Springer international series on information retrieval, vol 32. Springer, Berlin
- Müller H, Kalpathy-Cramer J, Eggel I, Bedrick S, Reisetter J, Kahn CE Jr, Hersh W (2010b) Overview of the CLEF 2010 medical image retrieval track. In: *Working notes of CLEF 2010 (Cross language evaluation forum)*
- Müller H, García Seco de Herrera A, Kalpathy-Cramer J, Demner Fushman D, Antani S, Eggel I (2012) Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In: *Working notes of CLEF 2012 (Cross language evaluation forum)*
- Radhouani S, Kalpathy-Cramer J, Bedrick S, Bakke B, Hersh W (2009) Multimodal medical image retrieval improving precision at ImageCLEF 2009. In: *Working notes of the 2009 CLEF workshop*, Corfu
- Rowe BR, Wood DW, Link AN, Simoni DA (2010) Economic impact assessment of NIST text retrieval conference (TREC) program. Technical report project number 0211875, National Institute of Standards and Technology

- Stefan LD, Ionescu B, Müller H (2017) Generating captions for medical images with a deep learning multi-hypothesis approach: ImageCLEF 2017 caption task. In: CLEF2017 working notes, CEUR Workshop Proceedings. Dublin, CEUR-WS.org. <http://ceur-ws.org>
- Thornley CV, Johnson AC, Smeaton AF, Lee H (2011) The scholarly impact of TRECVID (2003–2009). *J Am Soc Inf Sci Technol* 62(4):613–627
- Tommasi T, Caputo B, Welter P, Güld M, Deserno TM (2010) Overview of the CLEF 2009 medical image annotation track. In: Peters C, Caputo B, Gonzalo J, Jones G, Kalpathy-Cramer J, Müller H, Tsirikas T (eds) *Multilingual information access evaluation II. Multimedia experiments. Lecture notes in computer science*, vol 6242. Springer, Berlin, pp 85–93
- Tsirikas T, García Seco de Herrera A, Müller H (2011) Assessing the scholarly impact of ImageCLEF. In: CLEF 2011. *Springer lecture notes in computer science (LNCS)*, pp 95–106
- Tsirikas T, Larsen B, Müller H, Endrullis S, Rahm E (2013) The scholarly impact of CLEF (2000–2009). In: *Information access evaluation. Multilinguality, multimodality, and visualization*. Springer, Berlin, pp 1–12



# Automatic Image Annotation at ImageCLEF



Josiah Wang, Andrew Gilbert, Bart Thomee, and Mauricio Villegas

**Abstract** Automatic image annotation is the task of automatically assigning some form of semantic label to images, such as words, phrases or sentences describing the objects, attributes, actions, and scenes depicted in the image. In this chapter, we present an overview of the various automatic image annotation tasks that were organized in conjunction with the ImageCLEF track at CLEF between 2009–2016. Throughout the 8 years, the image annotation tasks have evolved from annotating Flickr photos by learning from clean data to annotating web images by learning from large-scale noisy web data. The tasks are divided into three distinct phases, and this chapter will provide a discussion for each of these phases. We will also compare and contrast other related benchmarking challenges, and provide some insights into the future of automatic image annotation.

## 1 Introduction

Millions of pictures are taken and shared across the globe via the Internet every day. Users have access to a flood of images, making it a challenge to locate and organize the ones they care about from this vast ocean of available visual data. An avid music

---

J. Wang (✉)

Department of Computing, Imperial College London, London, UK  
e-mail: [jw@josiahwang.com](mailto:jw@josiahwang.com)

A. Gilbert

CVSSP, University of Surrey, Guildford, UK  
e-mail: [a.gilbert@surrey.ac.uk](mailto:a.gilbert@surrey.ac.uk)

B. Thomee

Google, San Bruno, CA, USA  
e-mail: [bthomee@google.com](mailto:bthomee@google.com)

M. Villegas

omni.us, Berlin, Germany  
e-mail: [mauricio@omnius.com](mailto:mauricio@omnius.com)

fan might want to search for pictures of her favourite band performing at that festival a few months ago. A passionate ‘foodie’ might be seeking inspiration for his next gourmet adventure. Concerned citizens living abroad might want to find photos of the aftermath of a horrific earthquake in their home country. These are just a few of the many examples of why there is a need to identify ways to automatically ‘make sense of’ and ‘explain’ images.

ImageCLEF is a track run as part of the CLEF campaign that focuses on the retrieval and annotation of images. Introduced in 2003, ImageCLEF is one of the largest and longest running labs in CLEF. Since the inception of ImageCLEF, various tasks have been organized to facilitate progress in the field of image annotation and retrieval. These tasks were aimed at developing systems that are able to associate images with text, for example by learning to ‘describe’ images with labels or sentences, or by retrieving relevant images given a textual query, where the images were accompanied by additional text that could serve as potential cues (e.g. web pages, image captions). In the first editions of ImageCLEF, the focus was on retrieving images relevant to given multilingual queries from a web collection, while from 2006 onwards image annotation tasks were also introduced (Clough et al. 2007). The image annotation tasks initially focused on recognizing concrete visual object categories (*books, chairs, etc.*), but evolved later to cover higher-level semantic concepts (*water, sunny, day, beach, etc.*). ImageCLEF also covered various other tasks, such as those from the medical domain (Müller et al. 2007, 2008, 2009), plant identification (Goëau et al. 2011), and the retrieval of ‘lifelogs’ (Dang-Nguyen et al. 2017).

In this chapter, we focus on discussing the *automatic image annotation tasks* that were organized during eight consecutive years from 2009 to 2016. In contrast to earlier editions, the tasks during this period emphasized learning automatic image annotation from large-scale image datasets (with associated texts). The image annotation tasks can be divided into three distinct groups:

- **Flickr photo annotation (2009–2012):** Concept annotation of photos by learning from clean, manually labeled pictures.
- **Scalable concept image annotation (2012–2014):** Annotation of general images by learning from noisy, large-scale web data.
- **Concept annotation, localization and sentence generation (2015–2016):** Extension of the scalable concept image annotation task, including new subtasks like localizing concept instances and generating sentential descriptions for the images.

## 1.1 Outline

This chapter is organized as follows. In Sect. 2, we summarize the various image annotation tasks held in conjunction with ImageCLEF between 2009 and 2016. These include the Flickr photo annotation task (Sect. 2.1), the scalable concept image annotation task (Sect. 2.2) and the concept annotation, localization and sen-

tence generation task (Sect. 2.3). In Sect. 3 we provide an overview of and compare other related image annotation challenges. Finally, Sect. 4 offers conclusions and discusses the future direction of automatic image annotation.

## 2 Automatic Image Annotation @ ImageCLEF over the Years

Automatic image annotation is the task of automatically assigning some form of text to a given image, in order to provide a human-understandable explanation of the image. Traditionally, the annotations were in the form of textual labels (words or phrases), which could cover different categories or concepts, such as concrete visual objects (*cat, train*), scenes (*beach, city*), amorphous background elements (*sky, grass*) or abstract concepts (*scary, serene*). As the field progressed, more detailed image annotation tasks were introduced to provide a more fine-grained description of the image, such as attributes (*red car, furry dog*), actions (*playing computer, riding horse*), object localization (identifying where exactly the concept is located in the image), sentential descriptions (*a man riding a bicycle on the street*), and even generic image captions (*I spotted these beautiful roses while on a hike this morning*).

In this section we provide an overview of how the image annotation tasks at ImageCLEF evolved over the years, based on the tasks organized from 2009 until 2016. These tasks can be divided into three phases: (1) annotating Flickr photos using clean annotations; (2) annotating web images using noisy, large-scale web data; and (3) localizing concepts in and generating sentence descriptions for web images. For each phase we describe the tasks and the motivations behind organizing them, and also provide an overview of the participation and results of the tasks.

### 2.1 Flickr Photo Annotation, Years 2009–2012

The release of the MIRFLICKR (Huiskes and Lew 2008; Huiskes et al. 2010) collection opened the door to learning automatic image annotation techniques from a large collection of labeled photos. The dataset overcame many of the issues that affected existing collections, namely the photos (1) were freely and legally usable due to having a Creative Commons license, (2) were included as part of the dataset rather than just being referenced by a link, and (3) were accompanied by a wealth of metadata and precomputed features. Researchers now had access to a total of one million photos that were taken all over the world and annotated with descriptive keywords and captions. The ImageCLEF photo annotation task transitioned to the

Which of the following concepts are clearly present in the picture below? Tick all that apply

Cat

Dog

Horse


Fish

Bird

Insect

Animal (other)

None of the above



[click to view image in larger size](#)

**Fig. 1** An example crowdsourcing scenario, showing an image and a list of concepts the worker can annotate the image with

MIRFLICKR collection in 2009 (Nowak and Dunker 2010), which it kept using until the last edition in 2012 (Thomee and Popescu 2012).<sup>1</sup>

### 2.1.1 Task Description

The objective of the image annotation task was for participants to devise methods that could accurately predict what was depicted in a photo. As annotations can be somewhat noisy, in particular because many of the keywords may make sense only to the photographer that assigned them, the task organizers formed a diverse set of semantic concepts instead that would act as ground truth for the task. Unlike other image annotation tasks at the time that solely focused on recognizing physical objects, the newly defined concepts offered much more diversity and included items such as natural elements (e.g. *fog*, *reflection*), scenery (e.g. *coast*, *cityscape*), people (e.g. *age*, *relationship*), and emotion (e.g. *euphoric*, *scary*). About 50 concepts were defined in 2009, while this grew to just under 100 in the following years. Crowdsourcing was used to assign these concepts to 25,000 MIRFLICKR photos (see Fig. 1), where the presence or absence of a concept was judged by multiple annotators (see Fig. 2).

<sup>1</sup>Dataset for 2012 available at <http://doi.org/10.5281/zenodo.1246795>.



| concepts:              | agreement: | vote: |
|------------------------|------------|-------|
| timeofday_day          | 1.00       | X     |
| flora_flower           | 0.50       |       |
| quantity_none          | 1.00       | X     |
| quality_infocus        | 0.67       | X     |
| quality_selectivefocus | 0.33       |       |
| style_circularwarp     | 0.20       |       |
| style_overlay          | 0.80       | X     |
| view_closeupmacro      | 1.00       | X     |
| view_indoor            | 1.00       | X     |
| setting_fooddrink      | 1.00       | X     |
| sentiment_happy        | 1.00       | X     |

**Fig. 2** The concepts associated with the example image for which at least one worker indicated they were present, as well as the relative agreement between the workers and the outcome of the majority vote

### 2.1.2 Participation and Results

About 18 teams participated and submitted results during each of the years the Flickr photo annotation task was held. The results were evaluated per concept—how well was its presence and absence detected across all photos—and per photo—how well were the presences and absences of all concepts detected in a photo. Due to advances in evaluation measures and new insights, different measures were used each year to evaluate the results.

*Per Concept Evaluation* In 2009 the Equal Error Rate and the Area Under Curve were used, while this changed to Average Precision in 2010, and to Mean interpolated Average Precision in 2011. In 2012 both the interpolated and non-interpolated variants of the Mean Average Precision and the Geometric Mean Average Precision were determined, as well as the micro and macro F1-measures. Each of these evaluation measures has different properties and offers a distinct view of the performance of an image annotation method. For instance, the GMAP specifically highlights improvements obtained on relatively difficult concepts, such that increasing the average precision of a concept from 0.05 to 0.10 has a larger impact in its contribution to the GMAP than increasing the average precision from 0.25 to 0.30.

*Per Photo Evaluation* In 2009 the so-called Ontology Score was used, which was a hierarchical measure that calculated a score based on the distance of two concepts in the hierarchy, such that methods that annotated photos with concepts ‘close’ to the ground truth would score higher than methods that produced concepts that were ‘far away’. The Ontology Score was extended with a new cost map based on Flickr metadata in 2010 to better capture how people defined semantic closeness. In 2011 the Semantic R-Precision measure was used, a variant of R-Precision that also incorporated Flickr metadata to determine the semantic relatedness of visual concepts in the case of misclassifications. While a concept hierarchy was still used in

2012, the evaluation measures no longer focused on it and were the same measures as were used to evaluate the concept annotation performance as mentioned above.

Considering that each year the concepts to detect and/or the sampling from the MIRFLICKR dataset changed, the results obtained by the participants across the various editions of the tasks cannot be directly compared. However, we can still observe general trends in terms of the direction in which the image annotation field was moving, and which techniques appeared to perform better than others.

The teams that appeared at the top of the ranking in 2009 used combinations of global features (e.g. color histograms) and local features (e.g. salient points), where particularly representing the image as a spatial pyramid proved successful in combination with using an SVM as classifier or to a lesser extent logistic regression; in contrast, approaches only using global features did not fare as well. In 2010 the participants tried out a diverse repertoire of techniques, yet still the combination of global and local features performed best overall. The following year multimodal approaches, where textual and visual information were jointly considered using early or late fusion techniques, led to performance improvements together with special treatment of the textual metadata, such as word stemming, stop word removal, and incorporation of semantic and emotional connotations. Many of the participants of the 2012 edition took note of what worked well in previous tasks and presented methods that fused global and local visual features with semantically enriched textual features.

For the consecutive years where comparable evaluation measures were used it appears that the image annotation performance did not greatly improve. Considering that the ground truth annotation quality did not seem to differ between experts (used in the 2009 task) and crowdworkers (used in the later tasks), the most likely explanation for this seeming stagnation is that the difficulty of the concepts increased over the years, which the organizers themselves also noted. Indeed, during the 2010 edition of the task the average annotation accuracy for the concepts introduced the year before was substantially higher than for the concepts that were newly introduced; this would thus suggest that performance improved over time for at least the ‘easier’ concepts.

## ***2.2 Scalable Concept Image Annotation, Years 2012–2014***

A new image annotation task was proposed in 2012 to specifically address the problem of scalability in the number of concepts for image annotation. The first two editions (Villegas and Paredes 2012b; Villegas et al. 2013) were organized as subtasks<sup>2,3</sup> under the umbrella of the general photo annotation and retrieval task,

---

<sup>2</sup><http://imageclef.org/2012/photo-web>.

<sup>3</sup><http://imageclef.org/2013/photo/annotation>.

while it became the main image annotation task<sup>4</sup> in 2014 (Villegas and Paredes 2014).

When this series of tasks first began the greatest achievements in automatic image annotation were characterized by the reliance on mostly clean data specifically labeled for image annotation, a fact that greatly limits the scalability of the developed approaches. Most approaches today still rely on clean annotations. The main goal of the scalable concept image annotation tasks was to be an incentive for research groups working in this area to develop approaches capable of scaling concept-wise without the requirement of large amounts of human effort. To this end, a dataset was created (Villegas and Paredes 2012a,b) that consisted of data crawled from the web, containing pairs of images and web pages in which these images appeared (see Sect. 2.2.1). The motivation was that not only the images can be cheaply gathered for practically any topic from the web, but also that the text appearing near the images in the web pages may be related to the image content. The objective of the task was thus to develop annotation systems that are able to use this kind of automatically obtained weakly supervised data, while not permitting them to use data specifically labeled for image annotation. This way, such annotation systems are able to annotate images with new concepts for which they only need to automatically gather relevant data rather than relying on the existence of a manually labeled large set of images.

### 2.2.1 WEBUPV Scalable Image Annotation Datasets

To create the dataset, 31 million images and their corresponding web pages were downloaded by querying the image search engines of Google, Bing and Yahoo! using all words from the Aspell English dictionary, while filtering out duplicate web pages, near-duplicate images, and message-based images (e.g. containing the text ‘image removed’). Subsets of this large dataset were then selected for the ImageCLEF tasks<sup>5</sup> using the set of concepts chosen for the task. Only those concepts were chosen from the list of query words for which at least one image and web page were retrieved, while excluding those that produced too many results. For further details on the creation of the dataset please refer to Villegas and Paredes (2012a,b).

**Textual Features** Four sets of textual features were extracted per image and released to the participants: (1) the list of words used to find the image when querying the search engines, along with the rank position for the search engine(s) the word was found on, (2) the web page(s) in which the image appeared, (3) the image

---

<sup>4</sup><http://imageclef.org/2014/annotation>.

<sup>5</sup>Datasets available at

2012: <http://doi.org/10.5281/zenodo.1038533>

2013: <http://doi.org/10.5281/zenodo.257722>

2014: <http://doi.org/10.5281/zenodo.259758>.

URLs as referenced in the web pages it appeared in, and (4) the features obtained from the text extracted near the position(s) of the image in each web page it appeared in. To extract the text near the image, the web page contents were first converted to valid XML, after which script and style elements were removed. The extracted text then contained all unique terms within 600 words from the image location, not including HTML tags and attributes. The terms were weighted depending on the number of appearances, their distances to the image and the type of Document Object Model (DOM) element (e.g. title, alt, etc.) they appeared in. The resulting features include for each image at most the 100 word-score pairs with the highest weights. In addition, the web page title was also extracted.

**Visual Features** The images were resized so that the width and height had at most 240 pixels, while preserving their aspect ratios. In addition to the raw image content, the following features were extracted from the image content and also released to the participants: (1) color histograms dividing the image in  $3 \times 3$  regions, (2) bag-of-words of randomly projected and quantized dense sampled local color histograms, (3) *GIST* (Oliva and Torralba 2001) and (4) *SIFT*, *C-SIFT*, *RGB-SIFT* and *OPPONENT-SIFT* (van de Sande et al. 2010).

### 2.2.2 Task Descriptions

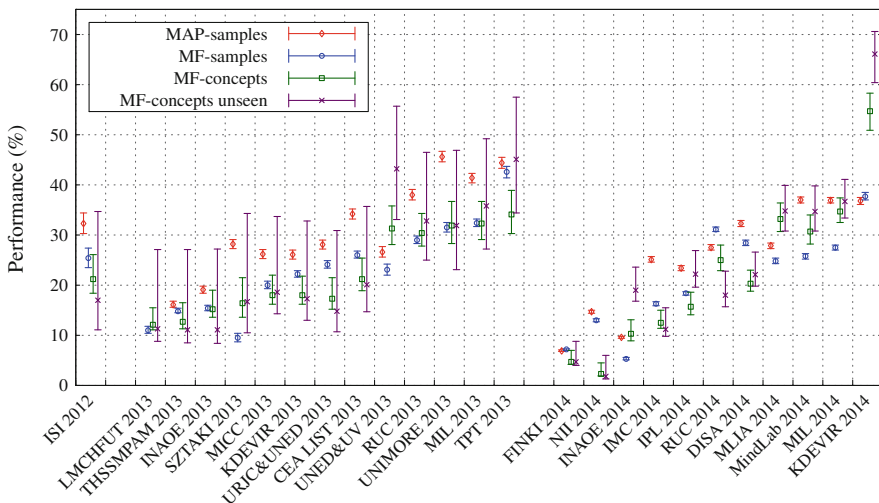
In 2012 two challenges were offered that participants could address. For each challenge a subset of 250,000 web images from the dataset was provided. In the first challenge the goal was to use the web data as a complement to the fully hand labeled data, whereby the list of concepts and test samples were exactly the same as the ones used in the Flickr annotation task of that year (see Sect. 2.1), thus making it possible to compare with the techniques that did not use the web data. In the second challenge the goal was to develop annotation systems using only automatically obtained data, i.e. the provided collection of images or something similar, as well as possibly using other language resources such as language models, language detectors, spell checkers, ontologies, automatic translation, etc. The use of any other labeled data was not permitted. The list of concepts was different for the development (95 concepts, 1000 images) and test (105 concepts, 2000 images) sets, and the participants only were able to use the ground truth annotations for the development set. The 2013 edition offered the second challenge again, whereby the 250,000 web images for training were the same, as were the development and test sets, although the ground truth had been refined (defined as WordNet synsets) and the concept list for the test set had been extended to 116 concepts. The 2014 edition significantly enlarged the dataset, whereby the training data of image web page pairs consisted of 500,000 images, half of which was the same as the previous two editions, while the development set consisted of 1940 samples labeled for 107 concepts. and the test set contained 7291 samples labeled for 207 concepts (100 unseen in development).



### 2.2.3 Participation and Results

The participation was very low in 2012, when only two groups took part in the subtask for which the web data was to be used to complement the manually labeled Flickr data, and only one group participated in the other web-only subtask. In contrast, the 2013 edition had a high participation, receiving 58 submissions from 13 different research groups. In 2014, 58 submissions were again received, but now from 11 different research groups. Both participants of the 2012 subtask on web data as complement actually obtained a worse annotation performance compared with using only the manually labeled data. Although much improvement could have been made in this direction, this subtask was not continued the following years as the main objective of the task was to work on scalable techniques.

Focusing on the web-only subtask, we show the results for all three editions in Fig. 3. The figure shows the test set results for the best submission of each participating group in each year, including the four performance measures that were used to analyze and judge the systems: mean average precision (MAP) for the samples and mean F-measure (MF) for the samples, the concepts and the subset of concepts that were not seen during development. The MF of unseen concepts is probably the most interesting since it shows how well the systems perform on new concepts. It can be observed that in 2013 the confidence intervals are wide, making it difficult to conclude what approaches performed best, and this was the main reason why the number of test samples and unseen concepts was enlarged significantly in 2014, which indeed resulted in much narrower confidence intervals.



**Fig. 3** Best submission of each participating group for the 2012, 2013 and 2014 editions. Error bars correspond to the 95% confidence intervals computed using Wilson's method

Diverse techniques were proposed by the participants, and the interested reader is encouraged to look at the details in the overview papers of each year and the working notes papers of the participants (Villegas and Paredes 2012b, 2014; Villegas et al. 2013). Here we will only mention a few ideas that proved successful. A general tendency was to use the web text data to select images for training, whereby it was common to make use of morphological expansions, stop words, word stemming and construction of ontologies using WordNet and Wikipedia. Another interesting idea introduced by TPT 2013 (Sahbi 2013), and also used by KDEVIR 2014 (Reshma et al. 2014), is the use of context dependent kernels for training SVMs, which are able to take advantage of the noisy textual data for training. To a certain extent the main objective of fostering the research of scalable annotation systems was achieved, and it was shown that it is possible to learn from such noisy data.

### ***2.3 Concept Annotation, Localization and Sentence Generation, Years 2015–2016***

There has been a big shift in the research landscape within the Computer Vision community since 2013. The successful performance of AlexNet (Krizhevsky et al. 2012) at the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC 2012) (Russakovsky et al. 2015) saw the rise of accurate Convolutional Neural Networks (CNN) based object classifiers trained on the large-scale ImageNet dataset, compared to the previously dominant methods using variants of SIFT-based bag-of-words representations. This improvement also resulted in the exploration of more complex annotation tasks beyond image-level labels, such as object localization and various joint vision and language tasks like image captioning.

To keep up with the change, the scalable concept image annotation task from 2012–2014 was revised in the 2015 edition (Gilbert et al. 2015) to make it more relevant and challenging. We retained the previous editions' goal of training automatic image annotation systems on noisy, large-scale (multimodal) web data rather than on cleanly annotated, concept-specific data. Unlike previous years, the use of CNNs trained on labelled ImageNet data was now allowed to accommodate the changes in methods in training state-of-the-art image classifiers. These CNN-based classifiers, however, were considered as baselines to encourage participants to come up with creative ways to utilise the noisy web data provided. Other changes and/or novelties introduced were:

- the extension of the classification task to include the localization of concepts;
- the introduction of a new sentence-level image description generation task;
- a larger, noisy web dataset of images with web pages ( $\approx 500\text{K}$ );
- a new list of 251 concepts for the annotation task;
- participants were asked to predict concepts/generate sentences for *all* images in the noisy training set; the test set was 'hidden' within the training set. This makes it a more challenging task in terms of scale.

As in the 2014 edition, the tasks in 2015 (Gilbert et al. 2015) and 2016 (Gilbert et al. 2016) were organized as main ImageCLEF tasks,<sup>6,7</sup> and were both divided into various subtasks.

### 2.3.1 Task Descriptions

As mentioned, the aim of the task was to evaluate different strategies to deal with noisy, large-scale data so that it can be reliably used for annotating, localizing, and generating natural sentences from practically any topic. Three subtasks were introduced in the 2015 and 2016 editions:

1. **Subtask 1** (*Image concept annotation and localization*): The image annotation task continued in the same line as the 2014 edition. The objective required the participants to develop a system that receives as input an image and produces as output a prediction of which concepts are present in that image, selected from a predefined list of concepts. In 2015–2016, participants were also asked to indicate where the concepts are located within the image (with a bounding box).
2. **Subtask 2** (*Image description generation*): Moving beyond annotating images with just concept labels, this subtask required the participants to describe images with a textual description of the visual content depicted in the image. It can be thought of as an extension of subtask 1, i.e. by generating full, sentential image descriptions from objects detected in subtask 1. This track was geared towards participants interested in developing systems that generated textual descriptions directly from images, e.g. by using visual detectors to identify concepts and generating textual descriptions from the detected concepts. This had a large overlap with subtask 1.
3. **Subtask 3** (*Image description generation/content selection from gold input*): This subtask is related to subtask 2, but aimed primarily at those interested only in the Natural Language Generation aspects of the subtask. For this subtask, a gold standard input (bounding boxes labelled with concepts) was provided to develop systems that generate sentential-based descriptions based on these gold standard annotations as input. In the 2015 edition, participants were asked to generate full sentence descriptions. In the 2016 edition, participants were only requested to provide a list of bounding box instances per image without having to generate the full description. This revision was so that evaluation can be focused on the *content selection* phase of image description generation, i.e. which concepts should be selected to be mentioned in the corresponding description?

Participants were allowed to take part in one or more subtasks. The subtasks were, however, designed in such a way that participants can take part in all three

---

<sup>6</sup><http://imageclef.org/2015/annotation>.

<sup>7</sup><http://imageclef.org/2016/annotation>.

subtasks as a pipeline, i.e. first detect and localize image concepts (subtask 1), and use these detections as input to generate image descriptions (subtasks 2 and 3) by reasoning about what should be described and how they should be described.

The 2016 edition also introduced a text illustration ‘teaser task’ to evaluate the performance of methods for text-to-image matching in news articles with images. We will not discuss this teaser task in this chapter. Instead, please refer to the 2016 task overview paper (Gilbert et al. 2016) for more details about the teaser task.

## Evaluation

To measure the performance of annotation and localization for subtask 1, the commonly used PASCAL VOC (Everingham et al. 2015) style metric of intersection over union (IoU) was used with respect to the ground truth. A *hard* and complex example of the ground truth for the object annotation and localization is shown in Fig. 4a.

The METEOR evaluation metric (Denkowski and Lavie 2014), adopted from the machine translation community, was used for subtask 2 and subtask 3 (2015 only). Commonly used to evaluate image description generation, METEOR is an *f*-measure-based measure that finds the optimal alignment of chunks of matched text, incorporating semantic knowledge by allowing terms to be matched to stemmed words, synonyms and paraphrases. Word ordering is accounted for by encouraging fewer matched chunks, indicating less fragmentation. METEOR matches a candidate text to each reference one-to-one, and takes the *maximum* score out of all references as the final score. METEOR has an advantage over other *n*-gram based metrics such as BLEU (Papineni et al. 2002) as it considers inexact word matches like synonyms and paraphrases from external resources.

The organizers also introduced a new fine-grained *content selection* metric for subtask 3 to evaluate how well the sentence generation system selects the correct concepts to be described against gold standard image descriptions (Gilbert et al. 2015; Wang and Gaizauskas 2015). The proposed metric essentially computes the *f*-measure against the gold standard concepts mentioned by humans; concepts that are more frequently mentioned are implicitly given a higher weight from the averaging process. An example of the ground truth annotation for subtasks 2 and 3 is shown in Fig. 4b.

### 2.3.2 Concepts

The 251 concepts for both 2015 and 2016 editions were chosen to be visual objects that are localizable and that are useful for generating textual descriptions of the visual content of images. They include animate objects such as people, dogs and cats, inanimate objects such as houses, cars and balls, and scenes such as city, sea and mountains. The concepts were mined from the texts of our dataset of 31 million image-webpage pairs (see Sect. 2.2.1 for more details). Nouns that are subjects or

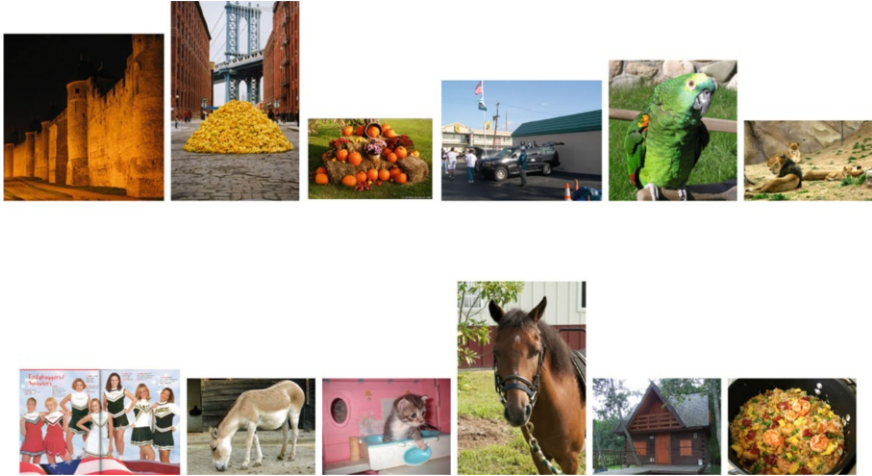


**Fig. 4** Examples of the ground truth for the image annotation and localization (subtask 1) and image description generation (subtasks 2 and 3). (a) Annotation example for subtask 1. (b) Image description generation example for subtask 2

objects of sentences were extracted and mapped onto WordNet synsets (Fellbaum 1998). These were then filtered to ‘natural’, basic-level categories (*dog* rather than a *Yorkshire terrier*), based on the WordNet hierarchy and heuristics from a large-scale text corpus (Wang et al. 2014). The organizers manually shortlisted the final list of 251 concepts such that they were (1) visually concrete and localizable; (2) suitable for use in image descriptions; (3) at a suitable ‘every day’ level of specificity that were neither too general nor too specific.

### 2.3.3 Dataset

As training dataset, the original WEBUPV Scalable Image Annotation database of 31 million images was used (Sect. 2.2.1). However, unlike 2012–2014 editions, a



**Fig. 5** Image examples from the test dataset

different subset was selected for the 2015 and 2016 editions to accommodate the new list of 251 concepts. More specifically, a subset of 500,000 images was selected from this database by choosing the top images from a ranked list of concepts and also a combination of concepts (so that an image contains more than one concept). The datasets including the test set ground truth are publicly available<sup>8</sup> for the research community.

Recognizing the increased computational power available, participants were expected to provide classification/localization (subtask 1) or to generate sentences (subtask 2) for *all* 500,000 ‘training’ images for the first time in the history of ImageCLEF image annotation tasks. Only a subset of test images ‘hidden’ within the ‘training’ set was used for evaluation purposes; these were not revealed to participants. Examples of the wide variety of the test set are shown in Fig. 5.

The development and test sets were both selected from the ‘training set’. The same test sets were used across both years and the results were as such comparable. A set of 5520 images was selected for the development and test data using a CNN trained to identify images suitable for sentence generation. 2000 of these were reserved as development data, while the remaining images were used as the test dataset. The images were then annotated via crowd-sourcing in three stages: (1) image level annotation for the 251 concepts; (2) bounding box annotation; (3) textual description annotation. An example of the crowd-sourcing interface is shown in Fig. 6. For subtask 3, the organizers further reserved 500 development and 450

<sup>8</sup>Datasets available at

2015: <http://doi.org/10.5281/zenodo.1038546>

2016: <http://doi.org/10.5281/zenodo.1038553>.

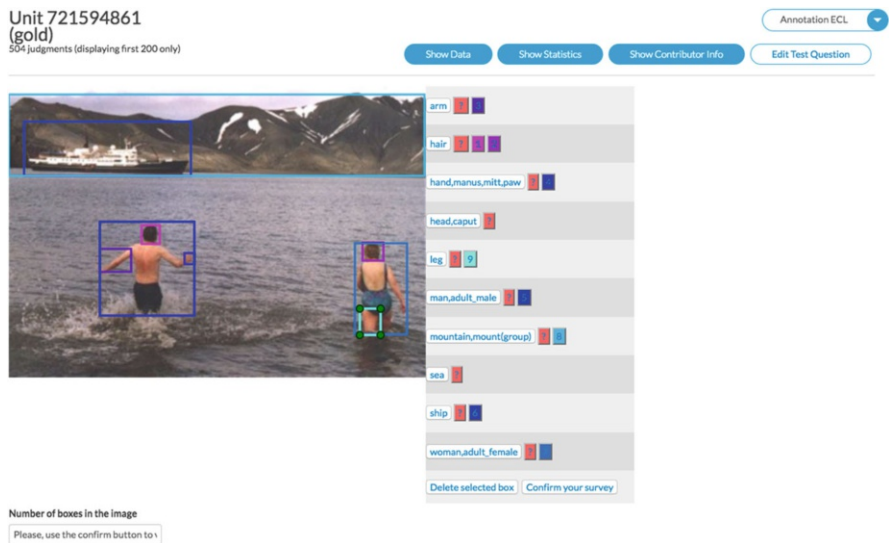


Fig. 6 The crowd sourcing interface to select the bounding box regions of interest

test images and annotated correspondences between bounding boxes and textual terms (Fig. 4b) for these to enable content selection evaluation.

Like the 2012–2014 editions, participants were provided with pre-computed textual and visual features, including *CNN* feature vectors extracted from the fully connected layer (fc7) of AlexNet (Krizhevsky et al. 2012) trained on the ILSVRC dataset (Russakovsky et al. 2015).

### 2.3.4 Participation and Results

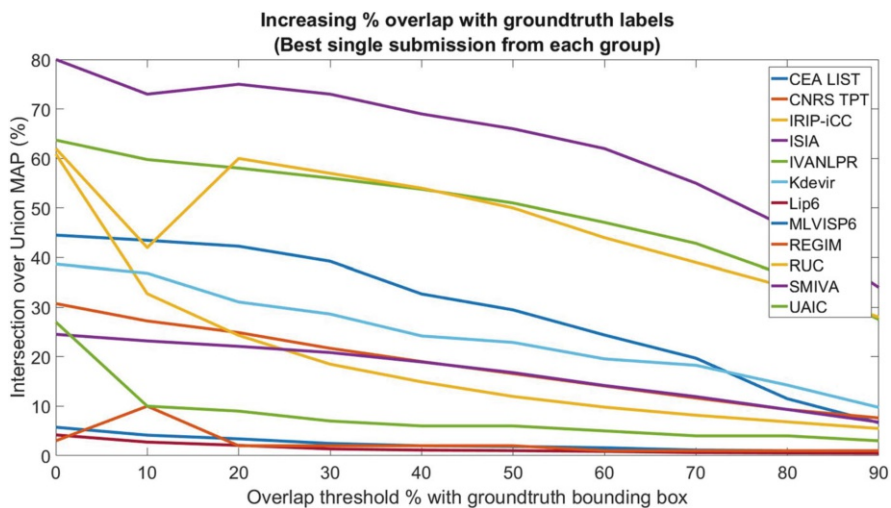
In both years good levels of participation were recorded, with 14 groups in 2015 and 7 in 2016, submitting over 150 and 50 runs respectively in the 2 years.

#### Subtask 1: Image Concept Annotation and Localization

Subtask 1 was well received despite the additional requirement of labeling and localizing all 500,000 images and the results for subtask 1 are presented in Table 1 in terms of mean average precision (MAP) over all images of all concepts, with both 0% overlap (i.e. no localization) and 50% overlap. A number of groups achieved excellent performance in 2015 with over 0.5 MAP. Both deep learning frameworks with additional annotated data, and SVM classifiers were used. The graph in Fig. 7 shows the performance of each submission for an increasing amount of overlap of the ground truth labels. In 2016 the method of computing the performance was

**Table 1** Subtask 1 results in 2015 and 2016

| Group               | 2015       |             | 2016       |             |
|---------------------|------------|-------------|------------|-------------|
|                     | 0% overlap | 50% overlap | 0% overlap | 50% overlap |
| SMIVA               | 0.79       | 0.66        |            |             |
| IVANLPR             | 0.64       | 0.51        |            |             |
| Multimedia comp lab | 0.62       | 0.50        |            |             |
| RUC                 | 0.61       | 0.50        |            |             |
| CEA                 | 0.45       | 0.29        | 0.54       | 0.37        |
| Kdevir              | 0.39       | 0.23        |            |             |
| ISIA                | 0.25       | 0.17        |            |             |
| CNRS-TPT            | 0.31       | 0.17        | 0.20       | 0.11        |
| IRIP-iCC            | 0.61       | 0.12        |            |             |
| UAIC                | 0.27       | 0.06        | 0.003      | 0.002       |
| MLVISP6             | 0.06       | 0.02        |            |             |
| REGIM               | 0.03       | 0.02        |            |             |
| Lip6                | 0.04       | 0.01        |            |             |
| MRIM                |            |             | 0.25       | 0.14        |



**Fig. 7** Increasing precision overlap of submissions for subtask 1

adjusted to include recall at a concept level, penalizing approaches that only detect a few concepts (for example face parts) by averaging the precision overall concepts. However, the approach by CEA, increased performance by around 8%, indicating continued progress. All approaches used a deep learning framework, including the Deep Residual Convolutional Neural Network (ResNet) (He et al. 2015). Face detection was fused into a number of approaches, however, in general, it was not found to provide much improvement in comparison to the improved neural network.



**Table 2** Results for subtask 2, showing the Meteor scores for the best run of all participants in both 2015 and 2016 editions

| Team           | Meteor              |
|----------------|---------------------|
| <i>Human</i>   | 0.3385 $\pm$ 0.1556 |
| ICTisia (2016) | 0.1837 $\pm$ 0.0847 |
| UAIC (2016)    | 0.0934 $\pm$ 0.0249 |
| RUC (2015)     | 0.1875 $\pm$ 0.0831 |
| ISIA (2015)    | 0.1644 $\pm$ 0.0842 |
| MindLab (2015) | 0.1403 $\pm$ 0.0564 |
| UAIC (2015)    | 0.0813 $\pm$ 0.0513 |

### Subtask 2: Image Description Generation

Subtask 2 had four participants in 2015 and two participants in 2016, with one team (UAIC) participating in both years. We observed a variety of approaches used to tackle these subtasks, including end-to-end neural models, template-based approaches, and joint image-text retrieval. Table 2 shows the results of the best run for all participants for both years. The best performing systems for both years utilized end-to-end CNN-LSTM image captioning systems, which at that point was the state of the art. The scores of the systems are, however, still significantly below the performance of the human upper-bound (by evaluating one description against the other descriptions for the same image and repeating the process for all descriptions). Thus, there is clear scope for future improvement and work to improve image description generation systems.

### Subtask 3: Image Description Generation/Content Selection from Gold Input

Two teams participated in subtask 3 for both 2015 and 2016 editions, again with one of the teams (UAIC) participating in both years. Table 3 shows the  $F$ -score, Precision and Recall across 450 test images for all participants of both years. RUC in 2015 chose to rerank the output of the state-of-the-art CNN-LSTM image captioning system using the provided gold input, which was a valid solution but not exactly what the organizers had envisioned. With the change in focus on evaluating only content selection in 2016, neither team that year relied on deep learning approaches for the task but instead concentrated on specifically solving the content selection task. Again, there is still scope for improvement for content selection compared to the estimated human upper-bound.

The 2015 and 2016 editions of the image annotation task successfully increased the ‘scalability’ aspect of image annotation, by doubling both the number of concepts and the size of the training set. The challenge for participants to annotate all 500,000 training images also did not deter participants from the tasks. New tasks were also introduced that go beyond annotating images with a single label and instead provide more meaningful and potentially more useful annotations with

**Table 3** Results for subtask 3, showing the content selection scores for the best run of all participants in both 2015 and 2016 editions

| Team            | Content selection score |                     |                     |
|-----------------|-------------------------|---------------------|---------------------|
|                 | Mean $F$                | Mean $P$            | Mean $R$            |
| <i>Human</i>    | $0.7445 \pm 0.1174$     | $0.7690 \pm 0.1090$ | $0.7690 \pm 0.1090$ |
| DUTh (2016)     | $0.5459 \pm 0.1533$     | $0.4451 \pm 0.1695$ | $0.7914 \pm 0.1960$ |
| UAIC (2016)     | $0.4982 \pm 0.1782$     | $0.4597 \pm 0.1553$ | $0.5951 \pm 0.2592$ |
| RUC (2015)      | $0.5310 \pm 0.2327$     | $0.6845 \pm 0.2999$ | $0.4771 \pm 0.2412$ |
| UAIC (2015)     | $0.5030 \pm 0.1775$     | $0.5095 \pm 0.1938$ | $0.5547 \pm 0.2415$ |
| <i>Baseline</i> | $0.1800 \pm 0.1973$     | $0.1983 \pm 0.2003$ | $0.1817 \pm 0.2227$ |

object localizations and sentential descriptions. While not many participants really utilized the large-scale, noisy web data in depth for training, there is still strong research potential in future for mining information from such a dataset.

### 3 Automatic Image Annotation: Beyond ImageCLEF

The ImageCLEF Image Annotation task was not the only benchmarking challenge in image annotation. Here we provide an overview of how the field has evolved over the years, and compare and contrast the main benchmarking tasks that were held in the same period.

**PASCAL Visual Object Classes (VOC)** This challenge was among the earliest image annotation tasks organized in the field of computer vision, which spanned the period 2005–2012 and had a focus on object classification, detection, and segmentation (Everingham et al. 2010, 2015). The accompanying dataset contained over 10,000 images with manually annotated bounding boxes and outlines for 20 object classes. In contrast with PASCAL VOC, the ImageCLEF Image Annotation tasks included concepts that went beyond mere objects and for which thus no bounding box or outline could necessarily be drawn; the recent editions also did away with cleanly annotated data to make learning the concepts more difficult and to be closer to how object recognition works in the real world where data is not perfect. This notwithstanding, PASCAL VOC has had a large influence in shaping and stimulating research on object recognition, and the recurring challenge made the yearly progress in recognition capabilities clearly visible.

**ImageNet Large Scale Visual Recognition Challenge (ILSVRC)** The improvements in classifier performance resulting from PASCAL VOC spurred the construction of larger image datasets such as ImageNet (Deng et al. 2009), which

included more than 20,000 object categories based on WordNet synsets, of which approximately 10,000 categories contained at least 100 images each. The release of the dataset led to the associated ILSVRC (Russakovsky et al. 2015) being organized, with the first edition held in 2010 alongside PASCAL VOC. The challenge scaled object recognition to 1000 categories formed by a subset of ImageNet synsets. It was also in the ILSVRC 2012 challenge that Convolutional Neural Networks (CNN) classifiers had their breakthrough, with AlexNet (Krizhevsky et al. 2012) achieving a significantly lower error rate compared to its competitors. After PASCAL VOC ended in 2012, the 2013 edition of ILSVRC introduced a new detection task that covered 200 categories. A principal objective of the challenge was for participants to correctly perform fine-grained recognition, such as for example different dog breeds, and unlike the ImageCLEF annotation tasks the concepts to be recognized were still related to low-level objects rather than also including higher-level semantics.

**Microsoft Common Objects in Context (COCO)** In contrast to the large number of categories of ILSVRC, the *Microsoft Common Objects in COntext (COCO)* dataset (Lin et al. 2014) had a different emphasis: recognizing a fewer number of categories (officially 80) but with substantially more examples per category. COCO aimed at the recognition of object instances in images in a ‘natural’ setting, where objects may be in the background or be visually occluded. The images therefore generally contained multiple object instances per image. Each image also was accompanied by at least five textual descriptions (Chen et al. 2015) to improve scene understanding. The creation of the dataset has resulted in various COCO benchmarking challenges, including object detection (bounding box and/or segmentation), image captioning, localization of person keypoints, and moving beyond concrete visual objects to ‘stuff’ segmentation (background regions like *sky* and *grass*) (Caesar et al. 2018). COCO has provided the research community with a great volume of richly annotated and segmented images; it differs from the ImageCLEF annotation tasks in that all of its objects still have to be visually identifiable and thus segmentable, whereas the ImageCLEF annotation tasks include numerous concepts that are not necessarily explicitly identifiable, such as *scary*, or can take on multiple shapes and forms, such as *motion blur*.

**MediaEval** This initiative encompasses a large number of benchmarks spanning audio, video, and images. MediaEval specifically serves as a platform for bringing together researchers working on problems in both relatively niche and well-established areas, where anyone can propose to organize a new task. The benchmark has organized several tasks in the past related to media annotation, such as predicting where in the world a photo or video was taken, which social events are depicted, which category a video belongs to, and which disease or condition can be identified in medical imagery. All these tasks though have very specific objectives in mind and a limited number of dimensions along which to annotate, in contrast with the ImageCLEF, PASCAL VOC, ILSVRC, and COCO challenges where the concepts to annotate with are generally more numerous and broader in scope.

In summary, all the mentioned challenges share a similar goal as the ImageCLEF image annotation tasks in improving image understanding. The ImageCLEF tasks, however, have focused on a larger variety of semantic concepts beyond visual objects, whereby also the use of large-scale, noisy web data compared to carefully labeled datasets was encouraged. As far as we are aware, no other challenges at this time have required participants to annotate  $\approx 500\text{K}$  images at test time.

## 4 Conclusion

In this chapter, we presented the automatic image annotation challenges organized as part of ImageCLEF between 2009–2016. The tasks evolved from annotating Flickr images using clean labels to annotating, localizing objects in, and describing large-scale web images using noisy web data. An overview of the tasks and results over the period was provided. We also compared the tasks to other image annotation tasks that were held during the same period and highlighted the fact that our tasks were different in that they focused on learning from large noisy data and with a larger variety of concepts.

While automatic image annotation traditionally involves annotating images with concepts based on pixel-level image understanding, the field has grown over the years, spurred by the availability of large-scale annotated data, powerful deep learning models, and computational resources to learn from such data. The field is moving towards a higher-level understanding of images, beyond just object or concept level, and has started leveraging image information for tasks like visual question answering (Antol et al. 2015) and visual dialogues (Das et al. 2017). It is hoped that we will eventually be able to build stronger systems that are able to achieve ‘general’ intelligence, by understanding images and text well enough for general tasks beyond task-specific datasets.

**Acknowledgements** The Concept Annotation, Localization and Sentence Generation task in ImageCLEF 2015 and 2016 were co-organized by the VisualSense (ViSen) consortium under the ERA-NET CHIST-ERA D2K 2011 Programme, jointly supported by UK EPSRC Grants EP/K019082/1 and EP/K01904X/1, French ANR Grant ANR-12-CHRI-0002-04 and Spanish MINECO Grant PCIN-2013-047.

## References

- Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D (2015) VQA: visual question answering. In: Proceedings of the IEEE international conference on computer vision (ICCV). IEEE, Piscataway, pp 2425–2433. <https://doi.org/10.1109/ICCV.2015.279>
- Caesar H, Uijlings J, Ferrari V (2018) COCO-stuff: thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR),

- pp 1209–1218. [http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Caesar\\_COCO-Stuff\\_Thing\\_and\\_CVPR\\_2018\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Caesar_COCO-Stuff_Thing_and_CVPR_2018_paper.html)
- Chen X, Fang H, Lin T, Vedantam R, Gupta S, Dollár P, Zitnick CL (2015) Microsoft COCO captions: data collection and evaluation server. CoRR abs/1504.00325. <http://arxiv.org/abs/1504.00325>. 1504.00325
- Clough P, Grubinger M, Deselaers T, Hanbury A, Müller H (2007) Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In: Peters C, Clough P, Gey FC, Karlgren J, Magnini B, Oard DW, de Rijke M, Stempfhuber M (eds) Evaluation of multilingual and multi-modal information retrieval: seventh workshop of the cross-language evaluation forum (CLEF 2006). Revised selected papers. Lecture notes in computer science (LNCS), vol 4730. Springer, Heidelberg, pp 223–256
- Dang-Nguyen DT, Piras L, Riegler M, Boato G, Zhou L, Gurrin C (2017) Overview of ImageCLEF2017: lifelog retrieval and summarization. In: Cappellato L, Ferro N, Goeuriot L, Mandl T (eds) CLEF 2017 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Das A, Kottur S, Gupta K, Singh A, Yadav D, Moura JMF, Parikh D, Batra D (2017) Visual dialog. In: Proceedings of the IEEE conference on computer vision and pattern recognition, IEEE, Piscataway, pp 1080–1089. <https://doi.org/10.1109/CVPR.2017.121>
- Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), IEEE, Piscataway, pp 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
- Denkowski M, Lavie A (2014) Meteor universal: language specific translation evaluation for any target language. In: Proceedings of the ninth workshop on statistical machine translation. Association for computational linguistics, pp 376–380. <https://doi.org/10.3115/v1/W14-3348>
- Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The PASCAL visual object classes (VOC) challenge. *Int J Comput Vis* 88(2):303–338. <https://doi.org/10.1007/s11263-009-0275-4>
- Everingham M, Eslami SMA, Van Gool L, Williams CKI, Winn J, Zisserman A (2015) The PASCAL visual object classes challenge: a retrospective. *Int J Comput Vis* 111(1):98–136. <https://doi.org/10.1007/s11263-014-0733-5>
- Fellbaum C (ed) (1998) WordNet an electronic lexical database. MIT Press, Cambridge
- Gilbert A, Piras L, Wang J, Yan F, Dellandrea E, Gaizauskas R, Villegas M, Mikolajczyk K (2015) Overview of the ImageCLEF 2015 scalable image annotation, localization and sentence generation task. In: Cappellato L, Ferro N, Jones GJF, SanJuan E (eds) CLEF 2015 labs and workshops, Notebook papers. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Gilbert A, Piras L, Wang J, Yan F, Ramisa A, Dellandrea E, Gaizauskas R, Villegas M, Mikolajczyk K (2016) Overview of the ImageCLEF 2016 scalable concept image annotation task. In: Balog K, Cappellato L, Ferro N, Macdonald C (eds) CLEF 2016 working notes. CEUR workshop proceedings (CEUR-WS.org), pp 254–278. ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Goëau H, Bonnet P, Joly A, Boujemaa N, Barthelemy D, Molino JF, Birnbaum P, Mouysset E, Picard M (2011) The CLEF 2011 plant images classification task. In: Petras V, Forner P, Clough P, Ferro N (eds) CLEF 2011 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1177/>
- He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 1026–1034. <https://doi.org/10.1109/ICCV.2015.123>
- Huiskes MJ, Lew MS (2008) The MIR flickr retrieval evaluation. In: Proceedings of the ACM international conference on multimedia information retrieval, pp 39–43
- Huiskes MJ, Thomee B, Lew MS (2010) New trends and ideas in visual concept detection: the MIR flickr retrieval evaluation initiative. In: Proceedings of the ACM international conference on multimedia information retrieval, pp 527–536

- Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) *Advances in neural information processing systems*, vol 25. Curran Associates, pp 1097–1105
- Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft COCO: common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) *Proceedings of the European conference on computer vision (ECCV)*. Springer, Berlin, pp 740–755
- Müller H, Deselaers T, Deserno TM, Clough P, Kim E, Hersh WR (2007) Overview of the ImageCLEFmed 2006 medical retrieval and medical annotation tasks. In: Peters C, Clough P, Gey FC, Karlgren J, Magnini B, Oard DW, de Rijke M, Stempfhuber M (eds) *Evaluation of multilingual and multi-modal information retrieval: seventh workshop of the cross-language evaluation forum (CLEF 2006)*. Revised selected papers. *Lecture notes in computer science (LNCS)*, vol 4730. Springer, Heidelberg, pp 595–608
- Müller H, Deselaers T, Deserno TM, Kalpathy-Cramer J, Kim E, Hersh WR (2008) Overview of the ImageCLEFmed 2007 medical retrieval and medical annotation tasks. In: Peters C, Jijkoun V, Mandl T, Müller H, Oard DW, Peñas A, Petras V, Santos D (eds) *Advances in multilingual and multimodal information retrieval: eighth workshop of the cross-language evaluation forum (CLEF 2007)*. Revised selected papers. *Lecture notes in computer science (LNCS)*, vol 5152. Springer, Heidelberg, pp 472–491
- Müller H, Kalpathy-Cramer J, Kahn CE, Hatt W, Bedrick S, Hersh W (2009) Overview of the ImageCLEFmed 2008 medical image retrieval task. In: Peters C, Deselaers T, Ferro N, Gonzalo J, Jones GJF, Kurimo M, Mandl T, Peñas A (eds) *Evaluating systems for multilingual and multimodal information access: ninth workshop of the cross-language evaluation forum (CLEF 2008)*. Revised selected papers. *Lecture notes in computer science (LNCS)*, vol 5706. Springer, Heidelberg, pp 512–522
- Nowak S, Dunker P (2010) Overview of the CLEF 2009 large-scale visual concept detection and annotation task. In: Peters C, Tsirikla T, Müller H, Kalpathy-Cramer J, Jones GJF, Gonzalo J, Caputo B (eds) *Multilingual information access evaluation vol. II multimedia experiments – tenth workshop of the cross-language evaluation forum (CLEF 2009)*. Revised selected papers. *Lecture notes in computer science (LNCS)*. Springer, Heidelberg, pp 94–109
- Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int J Comput Vis*. 42(3):145–175. <https://doi.org/10.1023/A:1011139631724>
- Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the association for computational linguistics (ACL)*, pp 311–318
- Reshma IA, Ullah MZ, Aono M (2014) KDEVIR at ImageCLEF 2014 scalable concept image annotation task: ontology based automatic image annotation. In: Cappellato L, Ferro N, Halvey M, Kraaij W (eds) *CLEF 2014 labs and workshops, Notebook papers, CEUR workshop proceedings (CEUR-WS.org)*. ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet large scale visual recognition challenge. *Int J Comput Vis* 115(3):211–252
- Sahbi H (2013) CNRS - TELECOM ParisTech at ImageCLEF 2013 scalable concept image annotation task: winning annotations with context dependent SVMs. In: Forner P, Navigli R, Tufis D, Ferro N (eds) *CLEF 2013 evaluation labs and workshop, Online working notes, CEUR workshop proceedings (CEUR-WS.org)*. ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Thomee B, Popescu A (2012) Overview of the ImageCLEF 2012 flickr photo annotation and retrieval task. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) *CLEF 2012 working notes. CEUR workshop proceedings (CEUR-WS.org)*. ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- van de Sande KE, Gevers T, Snoek CG (2010) Evaluating color descriptors for object and scene recognition. *IEEE Trans Pattern Anal Mach Intell* 32:1582–1596. <https://doi.org/10.1109/TPAMI.2009.154>

- Villegas M, Paredes R (2012a) Image-text dataset generation for image annotation and retrieval. In: Berlanga R, Rosso P (eds) II Congreso Español de Recuperación de Información, CERI 2012, Universidad Politécnica de Valencia, Valencia, pp 115–120
- Villegas M, Paredes R (2012b) Overview of the ImageCLEF 2012 scalable web image annotation task. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) CLEF 2012 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Villegas M, Paredes R (2014) Overview of the ImageCLEF 2014 scalable concept image annotation task. In: Cappellato L, Ferro N, Halvey M, Kraaij W (eds) CLEF 2014 labs and workshops, Notebook papers, CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>, pp 308–328
- Villegas M, Paredes R, Thomee B (2013) Overview of the ImageCLEF 2013 scalable concept image annotation subtask. In: Forner P, Navigli R, Tufis D, Ferro N (eds) CLEF 2013 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Wang J, Gaizauskas R (2015) Generating image descriptions with gold standard visual inputs: motivation, evaluation and baselines. In: Proceedings of the 15th European workshop on natural language generation (ENLG). Association for computational linguistics, pp 117–126
- Wang J, Yan F, Aker A, Gaizauskas R (2014) A poodle or a dog? Evaluating automatic image annotation using human descriptions at different levels of granularity. In: Proceedings of the third workshop on vision and language, Dublin City University and the association for computational linguistics, pp 38–45

# Image Retrieval Evaluation in Specific Domains



**Luca Piras, Barbara Caputo, Duc-Tien Dang-Nguyen, Michael Riegler, and Pål Halvorsen**

**Abstract** Image retrieval was, and still is, a hot topic in research. It comes with many challenges that changed over the years with the emergence of more advanced methods for analysis and enormous growth of images created, shared and consumed. This chapter gives an overview of domain-specific image retrieval evaluation approaches, which were part of the ImageCLEF evaluation campaign. Specifically, the robot vision, photo retrieval, scalable image annotation and lifelogging tasks are presented. The ImageCLEF medical activity is described in a separate chapter in this volume. Some of the presented tasks have been available for several years, whereas others are quite new (like lifelogging). This mix of new and old topics has been chosen to give the reader an idea about the development and trends within

---

L. Piras (✉)

University of Cagliari, Cagliari, Italy

Pluribus One, Cagliari, Italy

e-mail: [luca.piras@diee.unica.it](mailto:luca.piras@diee.unica.it); [luca.piras@pluribus-one.it](mailto:luca.piras@pluribus-one.it)

B. Caputo

University of Rome La Sapienza, Rome, Italy

e-mail: [caputo@diag.uniroma1.it](mailto:caputo@diag.uniroma1.it)

D.-T. Dang-Nguyen

University of Bergen, Bergen, Norway

Dublin City University, Dublin, Ireland

e-mail: [ductien.dangnguyen@uib.no](mailto:ductien.dangnguyen@uib.no)

M. Riegler

Simula Metropolitan Center for Digital Engineering and Kristiania University College, Oslo, Norway

e-mail: [michael@simula.no](mailto:michael@simula.no)

P. Halvorsen

Simula Metropolitan Center for Digital Engineering, Simula Research Laboratory and University of Oslo, Oslo, Norway

e-mail: [paalh@simula.no](mailto:paalh@simula.no)

© Springer Nature Switzerland AG 2019

N. Ferro, C. Peters (eds.), *Information Retrieval Evaluation in a Changing World*, The Information Retrieval Series 41, [https://doi.org/10.1007/978-3-030-22948-1\\_12](https://doi.org/10.1007/978-3-030-22948-1_12)

275



image retrieval. For each of the tasks, the datasets, participants, techniques used and lessons learned are presented and discussed leading to a comprehensive summary.

## 1 Introduction

In today's modern society, billions of people produce, upload and share digital images using devices like mobile phones or tablet computers. Autonomous intelligent machines like robots, drones and self-driving cars are equipped with RGB-D cameras, continuously capturing visual data to provide on-the-fly information about where the agent is, to subsequently inform its actions. Thus, images and more generally visual data have become a more important and natural way of communication ("a picture says more than thousand words"). This development is leading to large amounts of image data. Flickr, which started in 2004, reported until December 2017, a total number of around 6.5 billion uploaded photos, and Facebook reports around 300 million uploaded images per day. As one can see, there is an immanent need for methods supporting users with their image collections, and artificial machines with the understanding of their visual data. Retrieving images with a particular content, matching a particular query or for a particular purpose from large collections is challenging and has been a focus of research for much time. Likewise, recognizing objects, landmarks and scenes regardless of the imaging conditions are among the holy grails of computer vision since its infancy.

The ImageCLEF initiative and its community became aware of these emerging challenges at an early stage, and Image annotation and retrieval tasks have been part of ImageCLEF since 2003 (Clough and Sanderson 2004), while the Robot vision challenge was added in 2010 (Pronobis et al. 2010b). In this chapter, we provide an overview of the tasks over the years and how they developed. This gives a unique insight into how image retrieval and robot visual scene understanding changed over the years and which challenges emerged on the road.

In the early years, the focus was on retrieving relevant images from a web-collection given (multi-lingual) queries (Clough et al. 2005, 2006). From 2006 onwards, annotation tasks were also held, initially aimed at object detection (Clough et al. 2007; Deselaers et al. 2008), and more recently, also covering semantic concepts (Deselaers and Hanbury 2009; Nowak and Dunker 2010; Nowak and Huiskes 2010; Nowak et al. 2011; Thomee and Popescu 2012; Villegas and Paredes 2012, 2014; Villegas et al. 2013), photo retrieval (Grubinger et al. 2008; Arni et al. 2009; Lestari Paramita et al. 2010; Zellhöfer 2012, 2013), and robot vision (Pronobis et al. 2010a,b; Martínez-Gómez et al. 2012, 2013, 2014). In the last editions (Gilbert et al. 2015, 2016), the image annotation task was expanded to concept localization and also natural language sentential description of images. In recent years, there has been an increased interest in research combining text and vision. Therefore, in 2017, there has been a slight change in the focus of the retrieval goal. The task aimed at further stimulating and encouraging multi-modal research that uses text and visual data, and natural language processing for image retrieval and summarization (Dang-Nguyen et al. 2017b).

Although the tasks presented in this chapter are very diverse, spanning a large range of communities traditionally separated, it is still possible to identify a few crucial, shared lessons learned across the tasks over the years:

- multi-modal analysis improves performance compared to single-modal, but multi-modal is rarely exploited by researchers,
- methods change over time (for example switch to deep neural networks), but methods alone cannot solve the challenge entirely, and
- the larger the datasets gets, the harder it gets for participants to process them, whereas on the other side, too small datasets are also not interesting since they are not very applicable for deep learning.

All in all, image retrieval and visual place understanding for robotics applications still holds a lot of open research challenges that go far beyond simple classification such as semantics, object and landmarks detection and recognition, intent and personalized archives.

In the following sections, a detailed overview of selected tasks in the recent years is given. For each task, the data, participants, methods used and lessons learned are presented and discussed.

## **2 Tasks, Data and Participation**

### ***2.1 Overview of the Robot Vision Task***

The robot vision challenge was first organized in 2009, as part of the ImageCLEF lab (Pronobis et al. 2010b). Since then, the challenge has been organized another four consecutive times (Pronobis et al. 2010a,b; Martínez-Gómez et al. 2012, 2013, 2014). The first challenge focused on image-based place categorization, namely how to determine from a single image which room the robot is in. In its initial editions, the task focused exclusively on the use of RGB images for place categorization, and the participants were asked to process each image individually. Over the 5 years of the competition, the challenge grew in complexity so to include multi-modal data, and also in terms of the specific classification tasks required of the participants.

A very strong drive behind the organization of the Robot Vision Challenge was the need to provide the robot vision community with a benchmark where to measure quantitatively progress in semantic localization over the years. Indeed, performing repeatable experiments which produce quantitative, comparable results is a major challenge in robotics for many reasons. To begin with, running experiments often requires expensive hardware. Historically, such hardware has been almost always custom built and standardized, and complete robot platforms started to emerge only recently. Moreover, executing experiments involving real robots is often very time consuming and can be a major engineering challenge. As a result, a large chunk of robotics research has been evaluated in simulation or on a very limited scale.

By offering standardized benchmarks and publicly available databases, the Robot Vision Challenge has provided a tool allowing for fair comparisons, simplification of the experimental process, and as a result, a boost for progress in the field of robot vision.

### 2.1.1 Datasets

The Robot Vision Challenge was initially conceived as a visual place recognition competition, and the vision component has remained very strong in all its editions. Still, over the years, other additional tasks have been included. Table 1 illustrates these changes.

Accordingly, several datasets have been created over the years. The first dataset used in the challenge was the KTH-IDOL2 database (Luo et al. 2007). It was acquired using a mobile robot platform in the indoor environment of The Computer Vision and Active Perception laboratory (CVAP) at The Royal Institute of Technology (KTH) in Stockholm, Sweden. Each training image was annotated with the topological location of the robot and its pose  $\langle x; y; \theta \rangle$ . Although the pose information was provided in the training data, participants were discouraged from using it in their final submission. The two editions of the competition that took place in 2010 were based on COLD-Stockholm, an extension of *COsy Localization Database (COLD)* (Pronobis and Caputo 2009). This dataset was generated using a pair of high-quality cameras for stereo vision inside the same environment, as for the KTH-IDOL2 dataset. The fourth edition of the challenge used images from the unreleased VIDA dataset (Martínez-Gómez et al. 2013). This dataset includes perspective and range images acquired with a Kinect camera at the Idiap Research Institute in Martigny, Switzerland. Depth information was provided in the form of

**Table 1** Task evolution in the robot vision challenge

|            |                      | 1st edition | 2nd edition | 3rd edition | 4th edition | 5th edition |
|------------|----------------------|-------------|-------------|-------------|-------------|-------------|
| Sources    |                      |             |             |             |             |             |
|            | Monocular images     | X           | –           | –           | X           | X           |
|            | Stereo images        | –           | X           | X           | –           | –           |
|            | Depth images         | –           | –           | –           | X           | –           |
|            | Point clouds         | –           | –           | –           | –           | X           |
|            | Semantic annotations | X           | X           | X           | X           | X           |
|            | Pose annotations     | X           | –           | –           | –           | –           |
| Objectives |                      |             |             |             |             |             |
|            | Two tasks            | X           | X           | X           | X           | –           |
|            | Unknown classes      | –           | X           | X           | –           | –           |
|            | Kidnapping           | –           | –           | –           | X           | –           |
|            | Object detection     | –           | –           | –           | –           | X           |

**Table 2** Number of classes and training, validation and test set size

| Task edition | Number of classes | Training images | Validation images | Test images |
|--------------|-------------------|-----------------|-------------------|-------------|
| 1st          | 5                 | 2899            | 2789              | 1690        |
| 2nd          | 9                 | 12684           | 4783              | 5102        |
| 3rd          | 10                | 4782            | 2069              | 2741        |
| 4th          | 9                 | 7112            | 0                 | 6468        |
| 5th          | 10                | 5263            | 1869              | 3515        |

depth images, with color codes used to represent different distances. Finally, the fifth edition of the competition used images from the unreleased dataset ViDRiLO: The Visual and Depth Robot Indoor Localization with Objects information Dataset. This dataset includes images of the environment and point cloud files (in PCD format) (Martínez-Gómez et al. 2014). Table 2 summarizes the number of classes, as well as the number of training, validation and test images in each edition of the competition. It is worth underlining that the second and third editions included an unknown class not imaged in the training/validation sequences.

### 2.1.2 Evaluation Measures

The Robot Vision Challenge has always been focused on two main tasks, focused on visual place recognition. In the first one (mandatory task), participants have to provide information about the location of the robot for each test image from the data sequence perceived by the robot separately, i.e. without making any use of the label assigned to the image acquired at time  $t$  to make any prediction about the label to be assigned to the image at the time  $t + 1$ . In the second, optional task, instead the temporal continuity of the sequence can be used to improve the final classification of all images. The fifth edition of the challenge also introduced an object recognition task. Note that visual place recognition and object recognition can be considered as two subproblems of semantic localization, where each location is described in terms of its semantic contents.

As evaluation measure, all the editions used a score that computed the performance of the participant submission. This score was always based on positive values for test images correctly classified and negative values for misclassified ones. We also allowed the possibility to not classify test images, resulting in a non-effect on the score. The maximum reachable scores for the mandatory task of each edition were 1690, 5102, 2741, 2445, and 7030, respectively. Regarding the optional task, the maximum scores were 1690, 5102, 2741, and 4079, respectively, for the first to fourth editions. The fifth edition of the task had no optional task.

**Table 3** Participation to the robot vision challenge over the years

| Participation           | 1st edition | 2nd edition | 3rd edition | 4th edition | 5th edition |
|-------------------------|-------------|-------------|-------------|-------------|-------------|
| Registered groups       | 19          | 28          | 71          | 43          | 39          |
| Participant groups      | 7           | 8           | 7           | 8           | 6           |
| Working notes submitted | 5           | 3           | 3           | 4           | 2           |

### 2.1.3 Participants and Submissions

For all editions of the Robot Vision Challenge, a large number of groups registered, but only a small percentage of them actually participated in the competition and submitted results (see Table 3). We see that the number of registered groups grew considerably over the years, starting with 19 groups at the first edition, having a peak of 71 registered groups in third edition, and somehow reaching a plateau of roughly 40 groups registered in last two edition. In contrast, the number of groups actually participating in the challenges has been more stable over the years, with an average of 7 groups submitting their runs for the actual challenge every year. The submitted working notes were even less, with the highest number of working notes submitted in the first edition of the task (5 working notes submitted), and the minimum number reached for the last edition (2 working notes submitted).

### 2.1.4 Techniques Used

Nineteen different groups registered in the first edition of the Robot Vision Challenge (Pronobis et al. 2010b), organized in 2009. For the mandatory task, a wide range of techniques were proposed for the image representation and classification steps. The best result, 793 points out of 1690, was obtained by the Idiap group using a multi-cue discriminative approach (Xing and Pronobis 2009). The visual cues considered by this group included global as well as local descriptors, and then an *Support Vector Machines (SVM)* was trained for each visual cue and a high-level cue integration scheme *Discriminative Accumulation Scheme (DAS)* (Nilsback and Caputo 2004) was used to combine the scores provided by the different SVMs.

For the optional task, the best result, 916.5 points, was obtained by the SIMD group (Martínez-Gómez et al. 2009) using a particle filter approach to estimate the position of the robot given the previous position.

The 2010@ICPR edition (Pronobis et al. 2010a) had a participation similar to the first edition. Amongst the several proposals for the mandatory task, the approach adopted by the CVG group (Fraundorfer et al. 2010) stood out for its full usage of the stereo images to reconstruct the 3D geometry of the rooms, a choice that allowed them to achieve the best score of 3824 points out of 5102. The optional task was once again won by the SIMD group (Martínez-Gómez et al. 2010), where they this time computed similarities among local features between test frames and a set of training candidate frames, which was selected by means of clustering techniques.

In addition, a sort of temporal smoothing using prior assigned labels was used to classify very uncertain test frames.

The third edition of the Robot Vision challenge (Agosti et al. 2010) required the competing algorithms to show higher generalization capabilities compared to the previous editions. For the second time in a row, the mandatory task was won by the CVG group, with an approach combining a weighted k-NN search using global features, with a geometric verification step (Saurer et al. 2010). This approach obtained a score of 677 points out of 2741. For the optional task, the approach proposed by the Idiap group (Fornoni et al. 2010) was to be the most effective. The proposed multi-cue system combined up to three different visual descriptors in a discriminative multiple-kernel SVM. A door detector was implemented for discovering the transition from one room to another, while a stability estimation algorithm was used to evaluate the stability of the classification process.

As mentioned above, the 2012 edition (Martínez-Gómez et al. 2013) of the task introduced range images obtained with a Microsoft Kinect sensor. The organizers proposed a baseline method for both the feature extraction and the classification steps. The group from the Universidad Tecnológica Nacional, Córdoba, Argentina (CIII UTN FRC) (Sánchez-Oro et al. 2013) was the winner for both the mandatory and optional tasks with a score of 2071 and, 3930 respectively. It is worth noting, that this group was the only one that used depth information in their system. The fifth edition (Martínez-Gómez et al. 2014) encouraged participants to use 3D information (point cloud files) with the inclusion of rooms completely imaged in the dark, while also introducing the identification of objects in the scene. The highest result, 6033.5 points, was obtained by the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China (MIAR ICT) (Xu et al. 2014). They proposed the use of Kernel descriptors for both visual and depth information, while PCA was applied for dimensionality reduction. They used SVM classifiers and managed object recognition and room classification separately. Actually, both problems were expected to be handled together, but none of the participants presented a proposal where the objects appearance (or lack) is used to classify the room.

### 2.1.5 Lessons Learned

As a general remark, we can point out that, in most editions, the best techniques have been proposed by those participants taking advantage of the introduced novelties. Namely, those proposals that ranked first in the second and third editions were based on the spatial geometry acquired from stereo images. The top performing approach of the fourth edition was the only proposal using range information, and a similar scenario was found in the 2013 edition.

In addition to the generation of solutions to the problem provided by each edition, the Robot Vision task has served for sharing techniques and knowledge between worldwide researchers. This experience has supported several robotic laboratories for generating their own place classifiers, but also for the development of novel approaches that have been successfully deployed in different environments.

Overall, the Robot Vision challenge has provided valuable resources including datasets, benchmarking techniques, and state-of-the-art solutions to the visual place classification problem. Moreover, it has contributed to the generation of a semantic localization researchers community.

## ***2.2 Overview of the Photo Retrieval Task***

The Photo Retrieval Task was held over five editions starting in 2007 with the ImageCLEFphoto 2007 Photographic Retrieval Task (Grubinger et al. 2008). The first edition was followed by two very successful years in 2008 (Grubinger et al. 2008) and 2009 (Arni et al. 2009). After this, there was no new edition of the task until 2012 (Zellhöfer 2012) and 2013 (Zellhöfer 2013). The main idea of the Photo Retrieval Task was to perform laboratory-style evaluation of visual information retrieval from generic photo collections. In the first three editions of the task, the focus was put on light annotations, text and visual features using generic photo datasets. In the 2012 and 2013 editions, the focus was changed to personalized photo collections following the same principals for annotation and used modalities. The change was made as a consequence of a discussion at ImageCLEF 2011. This seemed not very appealing to participants, and the task did not reach the same number of participants or submissions for these 2 years, i.e., it was discontinued after 2013. In the following subsections, an overview is provided for the tasks in terms of datasets, metrics used, participants, techniques used and lessons learned.

### **2.2.1 Datasets**

In this subsection, the datasets used are discussed and presented in detail. As mentioned before, the first three editions of the task focused on generic datasets, whereas the two last editions put personalized photo collections into the focus. An overview of all tasks and used datasets can be found in Table 4.

In the 2007 task (Grubinger et al. 2008), 20,000 images were included in the dataset. In addition to the 20,000 color images, the organizers also included several types of metadata. The metadata provided to the participants includes annotation in different languages, title, location, date and additional notes. In addition, participants were given 60 different query topics (structured statements of user needs). The image features provided with the data were rather simple including color histograms, Tamura texture histograms and thumbnails compared with Euclidean distance.

The 2008 task (Arni et al. 2009) changed the focus from just photo retrieval to diversity of the result set, but using the same dataset as in 2007, i.e., the experimental results should show diverse representations of the query. This focus was also continued in the 2009 version of the task, but with a much larger dataset

**Table 4** Details of tasks per year

| Year | Task type                          | Resource  | Images  |
|------|------------------------------------|---|---------|
| 2007 | Photographic ad-hoc retrieval task | IAPR TC-12 Benchmark dataset                              | 20,000  |
| 2008 | Diversity in photo retrieval       | IAPR TC-12 Benchmark dataset                              | 20,000  |
| 2009 | Diversity in photo retrieval       | Belga dataset   | 498,920 |
| 2012 | Personal photo retrieval subtask   | Pythia dataset: uncompressed photographs of 19 laypersons | 5555    |
| 2013 | Personal photo retrieval subtask   | Pythia dataset: uncompressed photographs of 19 laypersons | 5555    |

(Lestari Paramita et al. 2010). The dataset for the 2009 task contained almost half a million images (498,920) which was by itself a challenge for the participants.

For both the 2012 (Zellhöfer 2012) and 2013 (Zellhöfer 2013) tasks, the dataset changed again. This time the focus was on personal photo collections, and the Pythia dataset containing only 5555 images was provided. The provided images were uncompressed and taken from 19 different laypersons. The ground truth for the dataset was created using relevance judgments which are highly subjective. From this dataset, the participants could use the following combinations of the provided document data and metadata: visual features alone, visual features and metadata, visual features and browsing data, metadata alone, metadata and browsing data, browsing data alone and a combination of all modalities.

### 2.2.2 Evaluation Measures

For the given tasks, a set of mainly well-known information retrieval metrics were used. This included average precision, precision at rank and F1-measures as harmonic mean of precision and recall. The first 2 years the tasks mainly focused on precision at a specific rank. To measure diversity, cluster recall was introduced in the 2008 editions of the task which also was used in 2009, but for a different rank and without precision. After a break of 2 years (2010 and 2011), the new tasks used *normalized Discounted Cumulative Gain (nDCG)* for the evaluation in addition to precision at rank 20 nDCG, which is known to be able to reflect subjectivity and evaluate relevance feedback, was chosen to compensate for the high subjective relevance judgments used to create the dataset ground truth. For the 2013 task, precision at rank 20 was not reported, but instead different cut offs of nDCG (5, 10, 20, 30, 100) and mean average precision with a cut off of 100 were used. Table 5 gives an overview of all metrics used for the different tasks. As one can see, the inconsistency of metrics used for the tasks makes it hard to compare results and measure improvements over the different editions. Nevertheless, since the focus of the task was basically shifted three times (retrieval, diversity, personalized image collections) this does not play an important role.



**Table 5** Details of metrics per year

| Year | Metrics  |
|------|--|
| 2007 | Mean average precision, precision at rank 20, geometric mean average precision, binary preference  |
| 2008 | Precision at rank 20, cluster recall at rank 20, mean average precision, geometric mean average precision, binary preference, F1-measure |
| 2009 | Precision at rank 10, cluster recall at rank 10, F1-measure  |
| 2012 | Mean average precision (cut off 100), nDCG <sup>a</sup>  |
| 2013 | nDCG   |

<sup>a</sup>Discounted cumulative gain of trec\_eval v9 with standard discount settings

**Table 6** Details of participants and submissions per year

| Year             | Registered | Participated   | Submitted runs  |
|------------------|------------|----------------|-----------------|
| 2007             | 32         | 20             | 616             |
| 2008             | >24        | 24             | 1042            |
| 2009             | 44         | 19             | 84              |
| 2012 (subtask 1) | 64         | 3 <sup>a</sup> | 13 <sup>a</sup> |
| 2012 (subtask 2) | 64         | 2 <sup>a</sup> | 8 <sup>a</sup>  |
| 2013             | 10         | 7              | 26              |

<sup>a</sup>The paper reports that there are excerpts presented in the result tables

### 2.2.3 Participants and Submissions

In general, the Photo Retrieval Task was well received by the community. Especially, the early editions attracted a large number of participants and submissions. In Table 6, the numbers of registered, participated and submitted runs are depicted. Registered indicates the number of people that were interested. Participated shows the number of distinctive teams that submitted a solution whereas submitted runs indicates how many runs the participating teams submitted in total. As mentioned before, the task was very popular in the first 3 years with a peak of participants and submitted runs of 24 and 1042 in 2008, respectively. In 2009, only 84 runs were submitted. This was most probably due to the large number of images in the dataset which could be a barrier for teams with not sufficient hardware.

The 2012 and 2013 tasks did not have as many submissions and participants compared to previous years. An explanation for this could not really be found. One reason could be the focus on personalized photo collections and the very subjective ground truth which makes evaluation fuzzy and maybe less interesting to the participants. Furthermore, the number of images in the dataset was also small compared to previous years with 20,000 and 498,920 images.

## 2.2.4 Techniques Used

During different editions of the task, participants used a vast number of different techniques and methods. This is depicted in the number of submitted runs, since for each run, something had to be different compared to a previous one. Overall, 1789 different runs were submitted in total (most of them for the first three editions). In the following, a summary about the most important aspects throughout all runs is provided.

In 2007 (Grubinger et al. 2008), most participants used the available image annotations for their analysis. The groups submitted 312 bilingual (combination of more than one language) runs and 251 monolingual. Two hundred and eighty eight runs were concept based (textual), 276 runs combined text and visual information and only 52 runs used only the visual content features. Most of the runs were based on automatic methods, but a small number also relied on manual approaches (around 3%). The most used language was English followed by German. The visual content information was the third most used modality.

In 2008 (Arni et al. 2009), the main questions asked for this task were (1) Is it possible to promote diversity withing the top n results?, (2) Which approaches work best for achieving diversity?, (3) Does diversity reduce the number of relevant images in the top n results?, (4) Can text retrieval be used to achieve diverse results?, and (5) How does the performance compare between bilingual and multilingual annotations? The dataset provided was the same as in the 2007 version in terms of images and provided metadata. Overall, 1024 runs were submitted where most of the submissions used the image annotations with 404 runs only using text information. 605 runs used visual information in combination with concept based features. Only 33 runs were purely content based (visual). Comparing the results for mixed, text-only and content-only, the best performance was achieved with mixed (text and visual) followed by text-only and visual-only on the last place. Most of the participants used different re-ranking methods or clustering for the analysis. Apart from that, different ways of merging modalities were applied like combining by scores, etc.

For the 2009 version of the task (Lestari Paramita et al. 2010), the participants were asked to specify the query fields used in their search and the modality of the runs. Query fields were described as T (Title), CT (Cluster Title), CD (Cluster Description) and I (Image). The modality was described as TXT (text-based search only), IMG (content-based image search only) or TXT-IMG (both text and content-based image search). This year, the highest F1 score was different for each modality. A combination of T-CT-I had the highest score in TXT-IMG modality. In the TXT modality, a combination of T-I scored the highest, with T- CT-I following on the second place. However, since only one run used the T-I, it was not enough to provide a conclusion about the best run. Calculating the average F1 score regardless of diversity shows that the best runs are achieved using a combination of Title, Cluster Title and Image. Using all tags in the queries resulted in the worst performance.

None of the participants for the 2012 task (Zellhöfer 2012) used a combination of all modalities (Zellhöfer 2012). The participants relied on visual features alone,

metadata alone, visual features and metadata, or metadata and browsing data. Interestingly, only one group decided to exploit the browsing data instead of the provided metadata. Surprisingly, they could use this data successfully to solve subtask 1, but reached the last position at subtask 2. This result indicates that there is a particularly strong influence of metadata on the retrieval of events.

Finally, in 2013 (Zellhöfer 2013), an interesting result of the conducted experiment was that the two leading groups performed almost equally well where one group was relying on sophisticated techniques such as Fisher vectors and local features while the other group used global low-end features embedded in a logical query language. Given the fact, that local features are computationally more intensive than global features, one might further investigate the logical combination of global features in order to achieve comparable results at less computational costs.

### 2.2.5 Lessons Learned

Based on the information collected from 5 years of the Photo Retrieval Tasks, several lessons can be learned. The overall conclusions that could be observed in all editions of the task are:

- Multi-modal analysis always improves the performance compared to only text, metadata or visual.
- Diversity is a topic that generates a lot of interest and is seen as important by the community.
- Personalized photo collections are less interesting for the community compared to more generic collections.
- Bilingual retrieval performs nearly as well as monolingual.

A more detailed analysis of the outcome of each task is provided in the following. Comparing the different combinations of provided features showed that using monolingual text achieves the best results followed by bilingual and visual information, respectively. In the monolingual results, Spanish outperforms English and German. In the bilingual results, a combination of English and German achieves the best results. The differences between the different languages are quite small, and the main conclusion was that the query language does only play a small role for the retrieval results. Comparing mixed, text- and visual-only runs, the mixed results (visual + test) outperform the text or visual only results by around 24% on average. Another insight is that manual methods outperform automatic methods, but these are not scalable and therefore unrealistic to use (Grubinger et al. 2008).

The 2008 task holds the record of participants and submitted runs in the series of the photo retrieval task. This is an interesting indicator for the general focus over time on photo retrieval in the community which peaked in 2008 and then decreased. The participants experimented with all different modalities whereas text was still most commonly used. The main insights and lessons learned were that bilingual retrieval performs nearly as well as monolingual (which was also observed in previous years). Combining the concept- and content-based retrieval methods

leads to the best results, and the visual retrieval methods got more popular (Arni et al. 2009).

For the 2009 task (Lestari Paramita et al. 2010), the results showed that participants were able to present a diverse result without sacrificing precision. In addition, the results revealed the following insights:

- Information about the cluster title is essential for providing diverse results, as this enables participants to correctly present images based on each cluster. When the cluster information was not being used, the cluster recall score is proven to drop, which showed that participants need better approaches to predict the diversity.
- A combination of title, cluster title and image was proven to maximize the diversity and relevance of the search engine.
- Using mixed modality (text and image) in the runs managed to achieve the highest F1 compared to using only text or image features alone.

Considering the increasing interest of participants in ImageCLEFPhoto, the creation of the new collection was seen as a big achievement in that it provides a more realistic framework for the analysis of diversity and evaluation of retrieval systems aimed at promoting diverse results. The findings from this new collection were found to be promising, and we plan to make use of other diversity algorithms (Dang-Nguyen et al. 2017a) in the future to enable evaluation to be done more thoroughly.

Finally, from the 2012 (Zellhöfer 2012) and 2013 (Zellhöfer 2013) tasks, the following insights were gained:

- There was no interest in solving the so-called user-centered initiative of the subtasks. The initiative asked for an alternative representation of the top-k results offering a more diverse view onto the results to the user. This challenge reflects the assumption that a user-centered system should offer users good and varying retrieval results.
- Varying results are likely to compensate for the vagueness inherent in both retrieval and query formulation. Hence, an additional filtering or clustering of the result list could improve the effectiveness and efficiency (in terms of usability) of the retrieval process.
- It remains unclear, if this task was too complex or just out of the area of expertise of the participants that used the dataset for the first time.
- The best performing groups used visual low-level features and metadata to solve the task.
- Again, the utilization of multiple modalities can increase the retrieval effectiveness.

### ***2.3 Overview of the Scalable Image Annotation Tasks***

From 2012 to 2016 (Villegas and Paredes 2012, 2014; Villegas et al. 2013; Gilbert et al. 2015, 2016), ImageCLEF ran a Scalable Image Annotation task, to promote

research into the annotation and classification of images using large-scale and noisy web page data. The primary goal of the challenge was to encourage creative ideas of using web page data to improve image annotation and to develop techniques to allow computers to describe images reliably, localize different concepts depicted and generate descriptions of the scenes. In the 2015 edition (Gilbert et al. 2015), the image annotation task was expanded to concept localization and also natural language sentential description of images. In 2016 edition (Gilbert et al. 2016), the organizers further introduced a text illustration task, to evaluate systems that analyze a text document and select the best illustration for the text from a large collection of images provided.

The challenging issue that was the basis of the image annotation challenges is that every day, users face the ever-increasing quantity of data available to them trying to find the image on Google of their favorite actress, or the images of the news article someone mentioned at work. Although there are a huge number of images that can be cheaply found from the Internet and a significant amount of information about the image is present on the web pages, the relationship between the surrounding text and images varies greatly, with much of the text being redundant and unrelated. Despite the obvious benefits of using such information in automatic learning, the weak supervision it provides means that it remains a challenging problem.

### 2.3.1 Datasets

In the first edition (Villegas and Paredes 2012) of the task, the organizers proposed two subtasks. In the first, the scope was to use both automatically gathered Web data and labeled data to enhance the performance in comparison to using only the labeled data, and in the second, the focus was to use only automatically gathered Web data and language resources to develop a concept scalable annotation system. A training set with 250,000 unlabeled images and textual features and 15,000 images from Flickr, labeled for 94 concepts, was provided to the participants for the first subtask. For subtask 2, the participants were allowed to use only the 250,000 unlabeled images. The test set consisted of 10,000 labeled images for the same 94 concepts of the training set and 2000 labeled images for 105 concepts respectively. In 2013 and 2014 (Villegas et al. 2013; Villegas and Paredes 2014), only one task was proposed whose purpose was developing concept scalable image annotation systems using only automatically gathered Web data. In these editions participants were provided with 250,000 Web images and respective Web-pages in 2013 and 500,000 images and respective Web-pages in 2014. In the second edition, the development set was composed by 1000 labeled images for 95 concepts, but in the test set, there were 2,000 images, and the participants had to label them for 116 concepts. In 2014, the participants had to label 7291 samples for 207 concepts, 100 unseen in development (see Table 7). In 2015 and 2016 (Gilbert et al. 2015, 2016), the participants were provided with unlabeled Web images, their respective Web-pages, and textual features. In the fourth edition, two subtasks were proposed, the first was related to image annotation as usual adding also a localization requirement. In the second, a

**Table 7** Number of concepts and training, development and test set

| Task edition         | Number of concepts | Training images  | Development images | Test images |
|----------------------|--------------------|------------------|--------------------|-------------|
| 2012 (subtask 1)     | 94                 | 250,000 + 15,000 | –                  | 10,000      |
| 2012 (subtask 2)     | 105                | 250,000          | 1000               | 2000        |
| 2013                 | 116                | 250,000          | 1000               | 2000        |
| 2014                 | 207                | 500,000          | 1940               | 7291        |
| 2015                 | 251                | 500,000          | 1979               | 3070        |
| 2016                 | 251                | 510,000          | 2000               | 3070        |
| 2016 ('teaser' task) | 251                | 310,000          | 3000               | 200,000     |

completely new task, the participants were requested to develop a system that could describe an image with a textual description of the visual content depicted in the image. The development set contained 1979 and the test set 3070 labeled images. In a second track (“clean track”) of the second subtask participants were provided with a test set of 450 images with bounding boxes labeled with concepts. Both development set and test sets were subsets of the 500,000 training images. In 2016, the three subtasks remained the same (the “clean track” became a subtask) but the organizers increased the number of images in the training set (510,000) and in the development (2000). In addition, a “teaser” task was proposed where participants were asked to analyze a given text document and find the best illustration for it from a set of all available images. The training set consisted of approximately 300,000 documents from the entire corpus. The remaining 200,000 have been used for testing. A separate development set of about 3000 image-webpage pairs were also provided as a validation set for parameter tuning and optimization purposes.

### 2.3.2 Evaluation Measures

The performance measures have been computed for each of the test set images, and as a global measure, the mean of these measures has been obtained. The measures used from 2012 to 2014 for comparing the submitted systems were *Average Precision (AP)* and *F-measure*. In addition in 2012, the *Interpolated Average Precision (IAP)* was also used.

$$AP = \frac{1}{C} \sum_{c=1}^C \frac{c}{rank(c)} \quad (1)$$

$$IAP = \frac{1}{C} \sum_{c=1}^C \max_{c' \geq c} \frac{c'}{rank(c')} \quad (2)$$

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (3)$$

where ‘C’ is the number of ground truth concepts for the image, ‘rank(c)’ is the rank position of the c-th ranked ground truth concept, and ‘precision’ and ‘recall’ are respectively the precision and recall for the annotation decisions. AP and IAP depended only on the confidence scores, while the F-measure only depended on the annotation decisions given by participants to the images.

Since the number of concepts per image was small and variable, the AP and IAP have been computed using the rank positions (precision =  $c/\text{rank}(c)$ ), i.e., using every possible value of recall, instead of using some fixed values of recall. After obtaining the means, these measures can be referred to as: *Mean Average Precision (MAP)*, *Mean Interpolated Average Precision (MIAP)* and *Mean F-measure (MF)*, respectively.

In 2015 and 2016, the localization of subtasks 1 have been evaluated using the PASCAL style metric of *Intersection over Union (IoU)*: the area of intersection between the foreground in the output segmentation and the foreground in the ground-truth segmentation, divided by the area of their union. The final results have been presented both in terms of average performance over all images of all concepts, and also per concept performance over all images. Subtask 2 has been evaluated using the Meteor<sup>1</sup> evaluation metric against a minimum of five human-authored textual descriptions. Systems participating in the “clean track” (called subtask 3 in 2016) have additionally had the option of being evaluated with a fine-grained metric, which was the average F-measure across all test images on how well the text generation system selected the correct concepts to be described (against the ground truth). In 2016, in addition, for the ‘teaser’ task, the test images have been ranked according to their distance to the query article. Recall at the k-th rank position (R@K) of the ground truth image have been used as performance metrics. Several values of k have been used, and participants were asked to submit the top 100 ranked images.

### 2.3.3 Participants and Submissions

The Scalable Image Annotation tasks, over the years, have had changing fortunes. After a somewhat weak start in 2012, where compared to 55 registered groups that signed the license agreement and therefore had access for downloading the datasets, only 26 runs were submitted for three groups on the two subtasks, the number of participants increased up to 2015. In 2013 and 2014, 58 runs were submitted for 13 and 11 groups, respectively, increasing to 122 runs submitted in 2015 for 14 groups from all parts of the world including China, France, Tunisia, Colombia, Japan, Romania. In 2016, unfortunately, the participation was not as good as in previous years. In total, 82 groups signed the license agreement, but only seven groups took part in the task and submitted 50 system runs overall (see Table 8).

---

<sup>1</sup><http://www.cs.cmu.edu/~alavie/METEOR/>.

**Table 8** Participation in the scalable image annotation tasks over the years

| Participation           | 2012 | 2013 | 2014 | 2015 | 2016 |
|-------------------------|------|------|------|------|------|
| Registered groups       | 55   | 104  | 43   | 154  | 82   |
| Participant groups      | 3    | 13   | 11   | 14   | 7    |
| Working notes submitted | 2    | 9    | 9    | 11   | 7    |

### 2.3.4 Techniques Used

The first attempts to participate in the Scalable Image Annotation challenge were based mainly on scalability (Ushiku et al. 2012b) using a combination of several SIFT features. For annotation, they used an online learning method *Passive-Aggressive with Averaged Pairwise Loss (PAAPL)* and labeled the Web-data using the appearance of concept words in the textual features.

The following years, the participation was much higher. Most of the submitted runs significantly outperformed the baseline system, but very large differences can be observed amongst the systems. In 2013, for both MAP and MF, the improvement was from below 10% to over 40%. An interesting detail to note is that for MAP there were several top performing systems. However, when comparing to the respective MF measures, the *CNRS TELECOM ParisTech (TPT)* submissions (Sahbi 2013) clearly outperform the rest. The key difference between these was the method for deciding which concepts were selected for a given image. This leads us to believe that many of the approaches could be improved greatly by changing that last step of their systems. Many of the participants have chosen to use the same scheme as the baseline system proposed by organizers for selecting the concepts, i.e., the top N and fixed for all images. The number of concepts per image was expected to be variable, thus making this strategy less than optimal. In contrast to usual image annotation evaluations with labeled training data, this challenge required facing different problems, such as handling the noisy data, textual processing and multi-label annotations. This permitted the participants to concentrate their efforts in different aspects. Several teams extracted their own visual features, for which they observed improvements with respect to the features provided by the organizers. For the textual processing, several different approaches were tried by the participants and some of these teams such as MIL (Hidaka et al. 2013), UNIMORE (Grana et al. 2013), CEA LIST (Borgne et al. 2013), and URJC&UNED (Sánchez-Oro et al. 2013) reported that as more information and additional resources are used the performance of the systems improved.

After the first appearance in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012 (Krizhevsky et al. 2012), *Convolutional Neural Network (CNN)* gained great popularity for its good performances on many classification tasks. Thus, it is no wonder that in 2014, three groups based their system on CNN pre-trained using ImageNet. Two of the teams, MIL (Kanehira et al. 2014) and MindLab (Vanegas et al. 2014), used the CNN output of an intermediate layer as



a visual feature. In Krizhevsky et al. (2012), it has been shown also that features extracted from the upper layers of the CNN can also serve as good descriptors for image retrieval. It implies that a CNN trained for a specific task has acquired generic representation of objects that will be useful for all sorts of visual recognition tasks (Babenko et al. 2014; Piras and Giacinto 2017). The third team that used CNN was MLIA (Xu et al. 2014), which employed the synsets predicted by the CNN to clean the concepts automatically assigned using the Web page data. In this case, the performance of the system could be greatly affected if the concepts for annotation differ significantly from the ones of ImageNet. As in previous years, most of the teams proposed approaches based on trained classifiers. In the case of the MIL team, the classifier is multi-label so each time the list of concepts to detect changed, the classifier had to be retrained. However, the PAAPL (Ushiku et al. 2012a) algorithm of MIL is designed with special consideration of scalability, so in their case, it does not seem an issue. Another approach that has been adopted was the use of one classifier per concept, trained one concept at a time using positive and negative samples. For scalability, the learning should be based on a selection of negative images so that this process is independent of how many concepts there are. Although many groups found it adequate, with respect to a multi-label classifier this might not be the optimal approach.

In the fourth edition of the challenge, the requirement of labeling and localizing all 500,000 images was added. Three groups achieved over 0.50 MAP across the evaluation set with 50% overlap with the ground-truth. This is an excellent result given the challenging nature of the images used and the wide range of concepts provided. Also in this edition, CNNs have been the masters and allowed for large improvements in performance. All of the top 4 groups used CNNs in their pipeline for the feature description. *Social Media and Internet Vision Analytics Lab (SMIVA)* (Kakar et al. 2015) used a deep learning framework with additional annotated data, while IVANLPR (Li et al. 2015b) implemented a two-stage process, initially classifying at an image level with an SVM classifier, and then applying deep learning feature classification to provide localization. RUC (Li et al. 2015a) trained per concept, an ensemble of linear SVMs trained by Negative Bootstrap using CNN features as image representation. Concept localization was achieved by classifying object proposals generated by Selective Search. The approach by CEA LIST (Gadeski et al. 2015) was very simple, they just use the CNN features in a small grid based approach for localization. In this edition, two other subtasks were proposed. For subtask 2, participants were asked to generate sentence-level textual descriptions for all 500,000 training images and the systems were evaluated on a subset of 3070 instances. RUC (Li et al. 2015a) used the state-of-the-art deep learning based *CNN Long Short-Term Memory Network (CNN-LSTM)* caption generation system, MindLab (Pellegrin et al. 2015) employed a joint image-text retrieval approach, and UAIC (Calfa et al. 2015) a template-based approach. Two of the teams that participated in this subtask participated also in the subtask 3 where participants were asked to generate textual descriptions for 450 test images based on labeled bounding box input. RUC (Li et al. 2015a) used a deep learning based

sentence generator coupled with re-ranking based on the bounding box input, while UAIC (Calfa et al. 2015) used a template-based generator.

In the 2016 edition, subtask 1 had a lower participation than in the previous year. However, there were some excellent results showing improvements over previous editions. CEA LIST (Borgne et al. 2016) used a deep learning framework (Simonyan and Zisserman 2014), but focused on improving the localization of the concepts. They attempted to use a face body part detector, boosted by previous year's results. MRIM-LIG (Portaz et al. 2016) also used a classical deep learning framework and the object localization (Uijlings et al. 2013), where an *a priori* set of bounding boxes were defined which were expected to contain a single concept each. CNRS (Sahbi 2016) focused on concept detection and used label enrichment to increase the training data quantity in conjunction with an SVM and VGG (Simonyan and Zisserman 2014) deep network.

### 2.3.5 Lessons Learned

The Scalable Image Annotation challenge had the objective of taking advantage of automatically gathered image and textual Web data for training, in order to develop more scalable image annotation systems. Even if in the subtask 1 of the first edition none of the participants were able to use the web-data to obtain a better performance than when using only manually labeled data, in the subtask 2, the results were somewhat positive. In some cases, the performance was even comparable to good annotation systems learned using manually labeled data. Over the years, some groups participated several times (e.g., MIL, KDEVIR, CEA LIST, TPT, INAOE), and in every edition, they were able to improve the results obtained in the previous one in particular improving more for the MF measures. This indicates a greater success in the developed techniques for choosing the final annotated concepts. In 2015, the requirement of labeling and localizing all 500,000 images was introduced. However, a limitation in the dataset has arisen: the difficulty of ensuring the ground truth has 100% of concepts labeled. This was especially problematic as the concepts selected include fine-grained categories such as eyes and hands that are generally small but occur frequently in the dataset. In addition, it was difficult for annotators to reach a consensus in annotating bounding boxes for less well-defined categories such as trees and field. Another interesting aspect of this challenge that has been going on for a long time is that the increased CNN usage as the feature representation had improved localization techniques and the performances have been progressively improved even under this point of view.

## 2.4 Overview of the Lifelog Tasks

The main goal of the Lifelog task is to advance the state-of-the-art research in lifelogging as an application of information retrieval. To do this, for each

edition, a standard dataset was provided and together with the dataset, and tasks were introduced. In the first edition, ImageCLEFlifelog2017 (Dang-Nguyen et al. 2017b), *Lifelog Retrieval Task (LRT)* and *Lifelog Summarization Task (LST)* were introduced, while in the second edition, ImageCLEFlifelog2018 (Dang-Nguyen et al. 2018), the LRT task was improved and renamed as *Activities of Daily Living (ADL)* understanding task. The details of these tasks are:

- *Lifelog Retrieval Task (LRT)*. In this task, the participants had to analyse the lifelog data and for several specific queries, return the correct answers. For example: “*In a Meeting: Find the moment(s) in which the user was in a meeting at work with 2 or more people. To be considered relevant, the moment must occur at meeting room and must contain at least two colleagues sitting around a table at the meeting. Meetings that occur outside of my place of work are not relevant.*”
- *Lifelog Summarization Task (LST)*. In this task, the participants had to analyse all the images and summarize them according to specific requirements. For instance: “*Shopping: Summarize the moment(s) in which user doing shopping. To be relevant, the user must clearly be inside a supermarket or shopping stores (includes book store, convenient store, pharmacy, etc). Passing by or otherwise seeing a supermarket are not considered relevant if the user does not enter the shop to go shopping. Blurred or out of focus images are not relevant. Images that are covered (mostly by the lifelogger’s arm) are not relevant.*” In this task, not only the relevance is considered, but participants are also asked to provide the diversification of the selected images with respect to the target scenario.
- *Activities of Daily Living (ADL)* understanding task: For this task, given a period of time, e.g., “*From 13 August to 16 August*” or “*Every Saturday*”, the participants should analyse the lifelog data and provide a summarisation based on the selected concepts (provided by the task organizers) of ADL and the environmental settings/contexts in which these activities take place. Some examples of ADL concepts: “*Commuting (to work or other common venue)*”, “*Travelling (to a destination other than work, home or some other common social event)*”, and contexts: “*In an office environment*”, “*In a home*”, “*In an open space*”. The summarisation should be described as the frequency and time spent for ADL concepts and total time for contexts concepts. For example: ADL: “*Eating/drinking: 6 times, 90 min*”, “*Travelling: 1 time, 60 min*”; context: “*In an office environment: 500 min*”, “*In a church: 30 min*”.

#### 2.4.1 Datasets

In the first edition, ImageCLEFlifelog2017, the dataset was developed based on the dataset in *NII Testbeds and Community for Information access Research (NTCIR)-12* (Gurrin et al. 2016). This dataset consists of data from three lifeloggers for a period of about one month each. The data contains 88,124 wearable camera images (about 1500–2500 images per day), an *eXtensible Markup Language (XML)* description of 130 associated semantic locations (e.g., Starbucks cafe, McDonalds

**Table 9** Statistics of the lifelog dataset

| Year                     | 2017           | 2018             |
|--------------------------|----------------|------------------|
| Number of days           | 90             | 50               |
| Size of the dataset (GB) | 18.18          | 18.85            |
| Number of images         | 88,124         | 88,440           |
| Number of locations      | 130            | 135              |
| Biometrics information   | No             | Yes              |
| Visual concepts          | Caffe concepts | Microsoft CV API |
| Personal annotation      | No             | Yes              |
| Music information        | No             | Yes              |
| Number of LRT topics     | 36             | –                |
| Number of LST topics     | 15             | 20               |
| Number of ADL topics     | –              | 20               |

restaurant, home, work) and the four physical activities, detected by the Moves app<sup>2</sup> installed in the lifeloggers’s phone: walking, cycling, running and transport of the lifeloggers at a granularity of 1 min. Together with the locations, activities and visual concepts are provided as the output of the Caffe CNN-based visual concept detector (Jia et al. 2014). This classifier provided labels and probabilities for 1000 objects in every image. The accuracy of the Caffe visual concept detector is variable and is representative of the current generation of off-the-shelf visual analytics tools.

In the second edition, ImageCLEFlifelog2018, the dataset was developed based on the dataset in NTCIR-13 (Gurrin et al. 2017). This dataset contains richer data with respect to the previous edition, where more biometrics information was added. The visual concept was also improved by using Microsoft Computer Vision API.<sup>3</sup>

Participants were provided with two different sets of topics: the development set (devset) for developing and training their methods and the test set (testset) for the final evaluation. A summary of the data collection is shown in Table 9.

#### 2.4.2 Evaluation Measures

For the LRT, evaluation metrics based on nDCG at different depths, i.e.,  $nDCG@N$ , were used. In this task,  $N$  was chosen from {5, 10}, depending on the topics. In the LST, classic metrics were deployed:

- Cluster Recall at  $X$  ( $CR@X$ )—a metric that assesses how many different clusters from the ground truth are represented among the top  $X$  results;

<sup>2</sup><http://moves-app.com/>.

<sup>3</sup><https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/>.

- Precision at  $X$  ( $P@X$ )—measures the number of relevant photos among the top  $X$  results;
- F1-measure at  $X$  ( $F1@X$ )—the harmonic mean of the previous two.

Various cut off points were considered, e.g.,  $X = 5, 10, 20, 30, 40, 50$ . The official ranking metrics this year was the **F1-measure@10** or images, which gives equal importance to diversity (via  $CR@10$ ) and relevance (via  $P@10$ ).

In the ADL understanding task, the score is computed as follows:

$$ADL_{score} = \frac{1}{2} \left( \max\left\{0, 1 - \frac{|n - n_{gt}|}{n_{gt}}\right\} + \max\left\{0, 1 - \frac{|m - m_{gt}|}{m_{gt}}\right\} \right)$$

where  $n, n_{gt}$  are the submitted and ground-truth values for how many times the events occurred, respectively, and  $m, m_{gt}$  are the submitted and ground-truth values for how long the events happened, respectively.

### 2.4.3 Participants and Submissions

In the first edition challenging task, eleven teams were registered, of which three teams took part in the task and submitted overall 19 runs. All three participating teams submitted a working paper describing their system.

In the second run, the number of participants was considerably higher compared to 2017 with 25 registered teams which submitted in total 41 runs: 29 (21 official, 8 additional) for LMRT and 12 (8 official, 4 additional) for ADLT, from 7 teams from Brunei, Taiwan, Vietnam, Greece-Spain, Tunisia, Romania, and a multi-nation team from Ireland, Italy, Austria, and Norway. The approaches employed ranged from fully automatic to fully manual, from using a single information source provided by the task to using all information as well as integrating additional resources, from traditional learning methods (e.g., SVMs) to deep learning and ad-hoc rules (Table 10).

### 2.4.4 Techniques Used

In the first edition, most of the teams proposed to only explore the visual or combine visual and textual information. In Molino et al. (2017), the authors proposed a three-step method as follow: As a first step, they filtered out images with very homogeneous colors and with a high blurriness. Then, the system ranked the

**Table 10** Details of participants and submissions per year of the lifelog task

| Year | Registered | Participated | Submitted runs |
|------|------------|--------------|----------------|
| 2017 | 11         | 3            | 19             |
| 2018 | 25         | 7            | 41             |

remaining images and clustered the top ranked images into a series of events using either k-means or a hierarchical tree. Finally, they selected, in an iterative manner, as many images per cluster as to fill a fixed size bucket (50 as required by the tasks). The study in Dogariu and Ionescu (2017) proposed an approach that combines textual and visual information in the process of selecting the best candidates for the task's requirements. The run that they submitted relied solely on the information provided by the organizers and no additional annotations or external data, nor feedback from the users had been employed. Additionally, a multi-stage approach has been used. The algorithm starts by analyzing the concept detector's output provided by the organizers and selecting for each image only the most probable concepts. From the list of the topics, each of them has then been parsed such that only relevant words have been kept and information regarding location, activity and the targeted user are extracted as well. The images that did not fit the topic requirements have been removed and this shortlist of images is then subject to a clustering step. Finally, the results are pruned with the help of similarity scores computed using WordNet's built-in similarity distance functions.

Learning from the drawbacks of the first edition, most of the teams participating in the second edition exploited multi-modal data by combining visual, text, location and other information to solve the tasks, which is different from the previous year when often only one type of data was analyzed. Furthermore, deep learning was exploited by many teams (Tran et al. 2018; Dogariu and Ionescu 2018; Abdallah et al. 2018). For example, in Dogariu and Ionescu (2018), CAMPUS-UPB team extracted the visual concepts using a CNN approach and then combined the extracted features with other information and clustered them using K-means and re-ranked using the concepts and queried topics. In the method proposed by the Regim Lab team (Abdallah et al. 2018), combinations of visual features, textual features and a combination of both by XQuery FLOWR, then fine tuned by CNN architectures were used. For the visual features fine tuned CNN architectures were utilized. Beside exploiting multi-modal data and deep learning, natural language processing was also considered. NLP-Lab (Tang et al. 2018) team reduced user involvement during the retrieval by using natural language processing. In this method, visual concepts were extracted from the images and combined with textual knowledge to get rid of the noise. For ADL, the images are ranked by time and frequency, whereas for LRT ranking is performed exploiting similarity between image concepts and user queries.

Different from the competitive teams, the task organizer team proposed only baseline approaches, with the purpose to serve as referent results for the participants. In the first edition (Zhou et al. 2017), they proposed multiple approaches, from fully automatic to fully manual paradigm. These approaches started by grouping similar moments together based on time and concepts. By applying this chronological-based segmentation, the problem of image retrieval turned into image segments retrieval. Starting from a topic query, it is transformed into small queries where each is asking for a single piece of information of concepts, location, activity, and time. The moments that matched all of those requirements are returned as the retrieval results. In order to remove the non-relevant images, a filtering step

is applied on the retrieved images, by removing blurred images and images that are mainly covered by a huge object or by the arms of the user. Finally, the images are diversified into clusters and the top images that close to center are selected for the summarization, which can be done automatically or using *Relevance Feedback (RF)*. In the second edition, the organizers proposed an improved version of the baseline search engine (Zhou et al. 2018), named LIFER, and based on that, an interactive lifelog search interface was built allowing users to solve both tasks in the competition.

### 2.4.5 Lessons Learned

We learned that in order to retrieve moments from lifelog data efficiently, we should exploit and combine multi-modal information, from visual, textual, location information to biometrics and the usage data from the lifeloggers devices. Furthermore, we learned that lifelogging is following the trend in data analytics, meaning that deep learning is being exploited in most of the methods. We also learned that there is still room for improvement, since the best results are coming from the fine-tuned queries, which means we need more advanced techniques for bridging the gap between an abstraction of human needs and the multi-modal data.

Regarding the participants, the significant improvement of the second edition compared to the first one shows how interesting and challenging lifelog data is and that it holds much research potential.

All in all, the task was quite successful for the first two runs, tacking into account that lifelogging is a rather new and not common field. The tasks helped to raise more awareness for lifelogging in general, but also to point at the potential research questions such as the previous mentioned multi-modal analysis, system aspects for efficiency, etc.

As next steps, the organizers do not plan to enrich the dataset but rather provide richer data and narrow down the applications of the challenges (e.g., extend to health-care application).

## 3 Discussion and Conclusions

From the descriptions of the different image retrieval tasks, some overall lessons and insights can be gained. Specifically the following insights are the most important:

- The consistent discrepancy between the registered groups and those eventually participating in the various challenges is a clear sign of interest in the data, and perhaps even more into the evaluation measures and experimental protocols developed over the years by all organizers. This is a testament to the ability of ImageCLEF of influencing and steering research in the community towards challenging goals.

- All tasks have significantly evolved over their lifetime, managing a fine balance between building a core community of participants that could leverage over prior experience in participating in the tasks, and continuously pushing the envelope in proposing new, cutting edge challenges supporting timely research in the respective field. This is witnessed by the popularity of the data and setup developed over the years.
- Lastly, the fundamental vision behind ImageCLEF has not been particularly affected by the deep learning tidal wave that did hit the community in the last years. On the contrary, ImageCLEF has continued thriving during this paradigm shift, and in several circumstances, it has been able to take advantage of it.

In conclusion, over its very long lifetime, ImageCLEF has been consistently a firm reference point in visual benchmarking and reproducibility, providing resources, promoting fundamental research questions and overall contributing strongly to the quest for intelligent seeing machines.

## References

- Abdallah FB, Feki G, Ezzarka M, Ammar AB, Amar CB (2018) Regim Lab Team at ImageCLEF-Flifelog LMRT Task 2018. In: Cappellato L, Ferro N, Nie JY, Soulier L (eds) CLEF 2018 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-2125/>
- Agosti M, Ferro N, Peters C, de Rijke M, Smeaton A (eds) (2010) multilingual and multimodal information access evaluation. In: Proceedings of the international conference of the cross-language evaluation forum (CLEF 2010). Lecture notes in computer science (LNCS), vol 6360. Springer, Heidelberg
- Arni T, Clough P, Sanderson M, Grubinger M (2009) Overview of the ImageCLEFphoto 2008 photographic retrieval task. In: Peters C, Deselaers T, Ferro N, Gonzalo J, Jones GJF, Kurimo M, Mandl T, Peñas A (eds) Evaluating systems for multilingual and multimodal information access: ninth workshop of the cross-language evaluation forum (CLEF 2008). Revised selected papers. Lecture notes in computer science (LNCS), vol 5706. Springer, Heidelberg, pp 500–511
- Babenko A, Slesarev A, Chigorin A, Lempitsky VS (2014) Neural codes for image retrieval. In: Fleet DJ, Pajdla T, Schiele B, Tuytelaars T (eds) Computer vision - ECCV 2014 - 13th European conference, Zurich, Switzerland, September 6–12, 2014, proceedings, part I. Lecture notes in computer science, vol 8689. Springer, Berlin, pp 584–599. <https://doi.org/10.1007/978-3-319-10590-1>
- Balog K, Cappellato L, Ferro N, Macdonald C (eds) (2016) CLEF 2016 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Borgne HL, Popescu A, Znaidia A (2013) CEA list@ImageCLEF 2013: scalable concept image annotation. In: Former P, Navigli R, Tufis D, Ferro N (eds) CLEF 2013 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Borgne HL, Gadeski E, Chami I, Tran TQN, Tamaazousti Y, Gínsca A, Popescu A (2016) Image annotation and two paths to text illustration. In: Balog K, Cappellato L, Ferro N, Macdonald C (eds) CLEF 2016 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Braschler M, Harman DK, Pianta E, Ferro N (eds) (2010) CLEF 2010 Working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1176/>



- Calfa A, Sillion D, Bursuc AC, Acatrinei CP, Lupu RI, Cozma AE, Padurariu C, Iftene A (2015) Using textual and visual processing in scalable concept image annotation challenge. In: Cappellato L, Ferro N, Jones GJF, SanJuan E (eds) CLEF 2015 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Cappellato L, Ferro N, Halvey M, Kraaij W (eds) (2014) CLEF 2014 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Cappellato L, Ferro N, Jones GJF, SanJuan E (eds) (2015) CLEF 2015 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Cappellato L, Ferro N, Goeuriot L, Mandl T (eds) (2017) CLEF 2017 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Cappellato L, Ferro N, Nie JY, Soulier L (eds) (2018) CLEF 2018 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-2125/>
- Clough P, Sanderson M (2004) The CLEF 2003 cross language image retrieval track. In: Peters C, Braschler M, Gonzalo J, Kluck M (eds) Comparative evaluation of multilingual information access systems: fourth workshop of the cross-language evaluation forum (CLEF 2003), Revised selected papers. Lecture notes in computer science (LNCS), vol 3237, Springer, Heidelberg, pp 581–593
- Clough P, Müller H, Sanderson M (2005) The CLEF 2004 cross-language image retrieval track. In: Peters C, Clough P, Gonzalo J, Jones GJF, Kluck M, Magnini B (eds) Multilingual information access for text, speech and images: fifth workshop of the cross-language evaluation forum (CLEF 2004), Revised selected papers. Lecture notes in computer science (LNCS), vol 3491. Springer, Heidelberg, pp 597–613
- Clough P, Müller H, Deselaers T, Grubinger M, Lehmann TM, Jensen J, Hersh W (2006) The CLEF 2005 cross-language image retrieval track. In: Peters C, Gey FC, Gonzalo J, Jones GJF, Kluck M, Magnini B, Müller H, de Rijke M (eds) Accessing multilingual information repositories: sixth workshop of the cross-language evaluation forum (CLEF 2005). Revised selected papers. Lecture notes in computer science (LNCS), vol 4022. Springer, Heidelberg, pp 535–557
- Clough P, Grubinger M, Deselaers T, Hanbury A, Müller H (2007) Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In: Peters C, Clough P, Gey FC, Karlgren J, Magnini B, Oard DW, de Rijke M, Stempfhuber M (eds) Evaluation of multilingual and multi-modal information retrieval: seventh workshop of the cross-language evaluation forum (CLEF 2006). Revised selected papers. Lecture notes in computer science (LNCS), vol 4730. Springer, Heidelberg, pp 223–256
- Dang-Nguyen DT, Piras L, Giacinto G, Boato G, Natale FGBD (2017a) Multimodal retrieval with diversification and relevance feedback for tourist attraction images. *ACM Trans Multimedia Comput Commun Appl* 13(4):49:1–49:24. <http://doi.acm.org/10.1145/3103613>
- Dang-Nguyen DT, Piras L, Riegler M, Boato G, Zhou L, Gurrin C (2017b) Overview of ImageCLEF lifelog 2017: lifelog retrieval and summarization. In: Cappellato L, Ferro N, Goeuriot L, Mandl T (eds) CLEF 2017 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Dang-Nguyen DT, Piras L, Riegler M, Zhou L, Lux M, Gurrin C (2018) Overview of ImageCLEF-Flifelog 2018: daily living understanding and lifelog moment retrieval. In: Cappellato L, Ferro N, Nie JY, Soulier L (eds) CLEF 2018 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-2125/>
- Deselaers T, Hanbury A (2009) The visual concept detection task in ImageCLEF 2008. In: Peters C, Deselaers T, Ferro N, Gonzalo J, Jones GJF, Kurimo M, Mandl T, Peñas A (eds) (2009) Evaluating systems for multilingual and multimodal information access: ninth workshop of the cross-language evaluation forum (CLEF 2008). Revised selected papers. Lecture notes in computer science (LNCS), vol 5706. Springer, Heidelberg, pp 531–538

- Deselaers T, Hanbury A, Viitaniemi V, Benczúr AA, Brendel M, Daróczy B, Escalante Balderas HJ, Gevers T, Hernández-Gracidias CA, Hoi SCH, Laaksonen J, Li M, Marín Castro HM, Ney H, Rui X, Sebe N, Stöttinger J, Wu L (2008) Overview of the ImageCLEF 2007 object retrieval task. In: Peters C, Jijkoun V, Mandl T, Müller H, Oard DW, Peñas A, Petras V, Santos D (eds) *Advances in multilingual and multimodal information retrieval: eighth workshop of the cross-language evaluation forum (CLEF 2007)*. Revised selected papers. Lecture notes in computer science (LNCS), vol 5152, Springer, Heidelberg, pp 445–471
- Dogariu M, Ionescu B (2017) A textual filtering of HOG-based hierarchical clustering of lifelog data. In: Cappellato L, Ferro N, Goeuriot L, Mandl T (eds) *CLEF 2017 working notes*. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Dogariu M, Ionescu B (2018) Multimedia lab @ CAMPUS at ImageCLEFlifelog 2018 lifelog moment retrieval. In: Cappellato L, Ferro N, Nie JY, Soulier L (eds) *CLEF 2018 working notes*. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-2125/>
- Fornier P, Karlgren J, Womser-Hacker C, Ferro N (eds) (2012) *CLEF 2012 working notes*. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Fornier P, Navigli R, Tufis D, Ferro N (eds) (2013) *CLEF 2013 working notes*. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Fornoni M, Martínez-Gómez J, Caputo B (2010) A multi cue discriminative approach to semantic place classification. In: Agosti M, Ferro N, Peters C, de Rijke M, Smeaton A (eds) *Multilingual and multimodal information access evaluation*. In: *Proceedings of the international conference of the cross-language evaluation forum (CLEF 2010)*. Lecture notes in computer science (LNCS), vol 6360. Springer, Heidelberg
- Fraundorfer F, Wu C, Pollefeys M (2010) Methods for combined monocular and stereo mobile robot localization. In: Ünyay D, Çataltepe Z, Aksoy S (eds) *Recognizing patterns in signals, speech, images and videos - ICPR 2010 contests*, Istanbul, Turkey, August 23–26, 2010, Contest reports. Lecture notes in computer science, vol 6388. Springer, Berlin, pp 180–189. [https://doi.org/10.1007/978-3-642-17711-8\\_19](https://doi.org/10.1007/978-3-642-17711-8_19)
- Gadeski E, Borgne HL, Popescu A (2015) CEA list's participation to the scalable concept image annotation task of imageclef 2015. In: Cappellato L, Ferro N, Jones GJF, SanJuan E (eds) *CLEF 2015 working notes*. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Gilbert A, Piras L, Wang J, Yan F, Dellandréa E, Gaizauskas RJ, Villegas M, Mikolajczyk K (2015) Overview of the ImageCLEF 2015 scalable image annotation, localization and sentence generation task. In: Cappellato L, Ferro N, Jones GJF, SanJuan E (eds) *CLEF 2015 working notes*. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Gilbert A, Piras L, Wang J, Yan F, Ramisa A, Dellandréa E, Gaizauskas RJ, Villegas M, Mikolajczyk K (2016) Overview of the ImageCLEF 2016 scalable concept image annotation task. In: Balog K, Cappellato L, Ferro N, Macdonald C (eds) *CLEF 2016 working notes*. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Grana C, Serra G, Manfredi M, Cucchiara R, Martoglia R, Mandreoli F (2013) UNIMORE at imageclef 2013: scalable concept image annotation. In: Fornier P, Navigli R, Tufis D, Ferro N (eds) *CLEF 2013 working notes*. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Grubinger M, Clough P, Hanbury A, Müller H (2008) Overview of the ImageCLEFphoto 2007 photographic retrieval task. In: Peters C, Jijkoun V, Mandl T, Müller H, Oard DW, Peñas A, Petras V, Santos D (eds) *Advances in multilingual and multimodal information retrieval: eighth workshop of the cross-language evaluation forum (CLEF 2007)*. Revised selected papers. Lecture notes in computer science (LNCS), vol 5152, Springer, Heidelberg, pp 433–444
- Gurrin C, Joho H, Hopfgartner F, Zhou L, Albatul R (2016) NTCIR lifelog: the first test collection for lifelog research. In: *Proceedings of the 12th NTCIR conference on evaluation of information access technologies*, pp 705–708

- Gurrin C, Joho H, Hopfgartner F, Zhou L, Gupta R, Albatal R, Dang-Nguyen DT (2017) Overview of NTCIR-13 Lifelog-2 task. In: Proceedings of the 13th NTCIR conference on evaluation of information access technologies
- Hidaka M, Gunji N, Harada T (2013) MIL at imageclef 2013: scalable system for image annotation. In: Forner P, Navigli R, Tufis D, Ferro N (eds) CLEF 2013 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22Nd ACM international conference on multimedia, MM'14. ACM, New York, pp 675–678. <http://doi.acm.org/10.1145/2647868.2654889>
- Kakar P, Wang X, Chia AY (2015) Automatic image annotation using weakly labelled web data. In: Cappellato L, Ferro N, Jones GJF, SanJuan E (eds) CLEF 2015 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Kanehira A, Hidaka M, Mukuta Y, Tsuchiya Y, Mano T, Harada T (2014) MIL at imageclef 2014: scalable system for image annotation. In: Cappellato L, Ferro N, Halvey M, Kraaij W (eds) CLEF 2014 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Bartlett PL, Pereira FCN, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems 25: 26th annual conference on neural information processing systems 2012. Proceedings of a meeting held December 3–6, 2012, Lake Tahoe, Nevada, pp 1106–1114. <http://papers.nips.cc/book/advances-in-neural-information-processing-systems-25-2012>
- Lestari Paramita M, Sanderson M, Clough P (2010) Diversity in photo retrieval: overview of the ImageCLEFPhoto task 2009. In: Peters C, Tsirikas T, Müller H, Kalpathy-Cramer J, Jones GJF, Gonzalo J, Caputo B (eds) Multilingual information access evaluation vol. II multimedia experiments – tenth workshop of the cross-language evaluation forum (CLEF 2009). Revised selected papers. Lecture notes in computer science (LNCS). Springer, Heidelberg, pp 45–59
- Li X, Jin Q, Liao S, Liang J, He X, Huo Y, Lan W, Xiao B, Lu Y, Xu J (2015a) RUC-Tencent at imageclef 2015: concept detection, localization and sentence generation. In: Cappellato L, Ferro N, Jones GJF, SanJuan E (eds) CLEF 2015 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Li Y, Liu J, Wang Y, Liu B, Fu J, Gao Y, Wu H, Song H, Ying P, Lu H (2015b) Hybrid learning framework for large-scale web image annotation and localization. In: Cappellato L, Ferro N, Jones GJF, SanJuan E (eds) CLEF 2015 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Luo J, Pronobis A, Caputo B, Jensfelt P (2007) Incremental learning for place recognition in dynamic environments. In: 2007 IEEE/RSJ international conference on intelligent robots and systems, October 29–November 2, 2007, Sheraton Hotel and Marina, San Diego, pp 721–728
- Martínez-Gómez J, Jiménez-Picazo A, García-Varea I (2009) A particle-filter based self-localization method using invariant features as visual information. In: Peters C, Ferro N (eds) Working notes for CLEF 2009 workshop co-located with the 13th European conference on digital libraries (ECDL 2009), Corfù, Greece, September 30–October 2, 2009, CEUR-WS.org, CEUR workshop proceedings, vol 1175. <http://ceur-ws.org/Vol-1175/CLEF2009wn-ImageCLEF-MartinezGomezEt2009.pdf>
- Martínez-Gómez J, Jiménez-Picazo A, Gámez JA, García-Varea I (2010) Combining image invariant features and clustering techniques for visual place classification. In: Únay D, Çataltepe Z, Aksoy S (eds) Recognizing patterns in signals, speech, images and videos - ICPR 2010 contests, Istanbul, Turkey, August 23–26, 2010, Contest reports. Lecture notes in computer science, vol 6388. Springer, Berlin, pp 200–209. [https://doi.org/10.1007/978-3-642-17711-8\\_21](https://doi.org/10.1007/978-3-642-17711-8_21)

- Martínez-Gómez J, García-Varea I, Caputo B (2012) Overview of the ImageCLEF 2012 robot vision task. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) CLEF 2012 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Martínez-Gómez J, García-Varea I, Cazorla M, Caputo B (2013) Overview of the ImageCLEF 2013 robot vision task. In: Forner P, Navigli R, Tufis D, Ferro N (eds) CLEF 2013 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Martínez-Gómez J, García-Varea I, Cazorla M, Morell V (2014) Overview of the ImageCLEF 2014 robot vision task. In: Cappellato L, Ferro N, Halvey M, Kraaij W (eds) CLEF 2014 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Molino AGD, Mandal B, Lin J, Lim JH, Subbaraju V, Chandrasekhar V (2017) VC-I2R@ImageCLEF2017: ensemble of deep learned features for lifelog video summarization. In: Cappellato L, Ferro N, Goeuriot L, Mandl T (eds) CLEF 2017 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Nilsback M, Caputo B (2004) Cue integration through discriminative accumulation. In: 2004 IEEE computer society conference on computer vision and pattern recognition (CVPR 2004), with CD-ROM, 27 June–2 July 2004, Washington. IEEE Computer Society, Washington, pp 578–585. <http://doi.ieeecomputersociety.org/10.1109/CVPR.2004.67>
- Nowak S, Dunker P (2010) Overview of the CLEF 2009 large-scale visual concept detection and annotation task. In: Peters C, Tsirikia T, Müller H, Kalpathy-Cramer J, Jones GJF, Gonzalo J, Caputo B (eds) Multilingual information access evaluation vol. II multimedia experiments – tenth workshop of the cross-language evaluation forum (CLEF 2009). Revised selected papers. Lecture notes in computer science (LNCS). Springer, Heidelberg, pp 94–109
- Nowak S, Huiskes MJ (2010) New strategies for image annotation: overview of the photo annotation task at ImageCLEF 2010. In: Braschler M, Harman DK, Pianta E, Ferro N (eds) CLEF 2010 Working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1176/>
- Nowak S, Nagel K, Liebetau J (2011) The CLEF 2011 photo annotation and concept-based retrieval tasks. In: Petras V, Forner P, Clough P, Ferro N (eds) CLEF 2011 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1177/>
- Pellegrin L, Vanegas JA, Ovalle JEA, Beltrán V, Escalante HJ, Montes-y-Gómez M, González FA (2015) INAOE-UNAL at imageclef 2015: scalable concept image annotation. In: Cappellato L, Ferro N, Jones GJF, SanJuan E (eds) CLEF 2015 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Peters C, Jijkoun V, Mandl T, Müller H, Oard DW, Peñas A, Petras V, Santos D (eds) (2008) Advances in multilingual and multimodal information retrieval: eighth workshop of the cross-language evaluation forum (CLEF 2007). Revised selected papers. Lecture notes in computer science (LNCS), vol 5152, Springer, Heidelberg
- Peters C, Deselaers T, Ferro N, Gonzalo J, Jones GJF, Kurimo M, Mandl T, Peñas A (eds) (2009) Evaluating systems for multilingual and multimodal information access: ninth workshop of the cross-language evaluation forum (CLEF 2008). Revised selected papers. Lecture notes in computer science (LNCS), vol 5706. Springer, Heidelberg
- Peters C, Tsirikia T, Müller H, Kalpathy-Cramer J, Jones GJF, Gonzalo J, Caputo B (eds) (2010) Multilingual information access evaluation vol. II multimedia experiments – tenth workshop of the cross-language evaluation forum (CLEF 2009). Revised selected papers. Lecture notes in computer science (LNCS). Springer, Heidelberg
- Piras L, Giacinto G (2017) Information fusion in content based image retrieval: a comprehensive overview. *Inf Fusion* 37:50–60
- Portaz M, Budnik M, Mulhem P, Poignant J (2016) MRIM-LIG at imageclef 2016 scalable concept image annotation task. In: Balog K, Cappellato L, Ferro N, Macdonald C (eds) CLEF 2016 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>

- Pronobis A, Caputo B (2009) COLD: the cosy localization database. *I J Robot Res* 28(5):588–594
- Pronobis A, Fornoni M, Christensen HI, Caputo B (2010a) The robot vision track at ImageCLEF 2010. In: Braschler M, Harman DK, Pianta E, Ferro N (eds) CLEF 2010 Working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1176/>
- Pronobis A, Xing L, Caputo B (2010b) Overview of the CLEF 2009 robot vision track. In: Peters C, Tsirikla T, Müller H, Kalpathy-Cramer J, Jones GJF, Gonzalo J, Caputo B (eds) Multilingual information access evaluation vol. II multimedia experiments – tenth workshop of the cross-language evaluation forum (CLEF 2009). Revised selected papers. Lecture notes in computer science (LNCS). Springer, Heidelberg, pp 110–119
- Sahbi H (2013) CNRS - TELECOM ParisTech at ImageCLEF 2013 scalable concept image annotation task: winning annotations with context dependent SVMs. In: Forner P, Navigli R, Tufis D, Ferro N (eds) CLEF 2013 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Sahbi H (2016) CNRS TELECOM paristech at imageclef 2016 scalable concept image annotation task: overcoming the scarcity of training data. In: Balog K, Cappellato L, Ferro N, Macdonald C (eds) CLEF 2016 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Sánchez-Oro J, Montalvo S, Montemayor AS, Pantrigo JJ, Duarte A, Fresno V, Martínez-Unanue R (2013) Urjc&unet at imageclef 2013 photo annotation task. In: Forner P, Navigli R, Tufis D, Ferro N (eds) CLEF 2013 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Saurer O, Fraundorfer F, Pollefeys M (2010) Visual localization using global visual features and vanishing points. In: Agosti M, Ferro N, Peters C, de Rijke M, Smeaton A (eds) multilingual and multimodal information access evaluation. In: Proceedings of the international conference of the cross-language evaluation forum (CLEF 2010). Lecture notes in computer science (LNCS), vol 6360. Springer, Heidelberg
- Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556. [1409.1556](https://arxiv.org/abs/1409.1556)
- Tang TH, Fu1 MH, Huang HH, Chen KT, Chen HH (2018) NTU NLP-Lab at ImageCLEF lifelog 2018: visual concept selection with textual knowledge for understanding activities of daily living and life moment retrieval. In: Cappellato L, Ferro N, Nie JY, Soulier L (eds) CLEF 2018 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-2125/>
- Thomee B, Popescu A (2012) Overview of the ImageCLEF 2012 flickr photo annotation and retrieval task. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) CLEF 2012 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Tran MT, Truong TD, Dinh-Duy T, Vo-Ho VK, Luong QA, Nguyen VT (2018) Lifelog moment retrieval with visual concept fusion and text-based query expansion. In: Cappellato L, Ferro N, Nie JY, Soulier L (eds) CLEF 2018 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-2125/>
- Uijlings JRR, van de Sande KEA, Gevers T, Smeulders AWM (2013) Selective search for object recognition. *Int J Comput Vis* 104(2):154–171
- Ünay D, Çataltepe Z, Aksoy S (eds) (2010) Recognizing patterns in signals, speech, images and videos - ICPR 2010 contests, Istanbul, Turkey, August 23–26, 2010, Contest reports. Lecture notes in computer science, vol 6388. Springer, Berlin. <https://doi.org/10.1007/978-3-642-17711-8>
- Ushiku Y, Harada T, Kuniyoshi Y (2012a) Efficient image annotation for automatic sentence generation. In: Babaguchi N, Aizawa K, Smith JR, Satoh S, Plagemann T, Hua X, Yan R (eds) Proceedings of the 20th ACM multimedia conference, MM'12, Nara, Japan, October 29–November 02, 2012. ACM, New York, pp 549–558

- Ushiku Y, Muraoka H, Inaba S, Fujisawa T, Yasumoto K, Gunji N, Higuchi T, Hara Y, Harada T, Kuniyoshi Y (2012b) ISI at imageclef 2012: scalable system for image annotation. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) CLEF 2012 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Vanegas JA, Ovalle JEA, Montenegro JSO, Páez F, Pérez-Rubiano SA, González FA (2014) Mindlab at imageclef 2014: scalable concept image annotation. In: Cappellato L, Ferro N, Halvey M, Kraaij W (eds) CLEF 2014 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Villegas M, Paredes R (2012) Overview of the ImageCLEF 2012 scalable web image annotation task. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) CLEF 2012 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Villegas M, Paredes R (2014) Overview of the ImageCLEF 2014 scalable concept image annotation task. In: Cappellato L, Ferro N, Halvey M, Kraaij W (eds) CLEF 2014 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Villegas M, Paredes R, Thomee B (2013) Overview of the ImageCLEF 2013 scalable concept image annotation subtask. In: Forner P, Navigli R, Tufis D, Ferro N (eds) CLEF 2013 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- King L, Pronobis A (2009) Multi-cue discriminative place recognition. In: Peters C, Tsirikla T, Müller H, Kalpathy-Cramer J, Jones GJF, Gonzalo J, Caputo B (eds) Multilingual information access evaluation vol. II multimedia experiments – tenth workshop of the cross-language evaluation forum (CLEF 2009). Revised selected papers. Lecture notes in computer science (LNCS). Springer, Heidelberg, pp 315–323
- Xu X, Shimada A, Taniguchi R (2014) MLIA at ImageCLEF 2014 scalable concept image annotation challenge. In: Cappellato L, Ferro N, Halvey M, Kraaij W (eds) CLEF 2014 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Zellhöfer D (2012) Overview of the personal photo retrieval pilot task at ImageCLEF 2012. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) CLEF 2012 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Zellhöfer D (2013) Overview of the ImageCLEF 2013 personal photo retrieval subtask. In: Forner P, Navigli R, Tufis D, Ferro N (eds) CLEF 2013 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Zhou L, Piras L, Riegler M, Boato G, Dang-Nguyen DT, Gurrin C (2017) Organizer team at ImageCLEFlifelog 2017: baseline approaches for lifelog retrieval and summarization. In: Cappellato L, Ferro N, Goeuriot L, Mandl T (eds) CLEF 2017 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Zhou L, Piras L, Riegler M, Lux M, Dang-Nguyen DT, Gurrin C (2018) An interactive lifelog retrieval system for activities of daily living understanding. In: Cappellato L, Ferro N, Nie JY, Soulier L (eds) CLEF 2018 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-2125/>

# About Sound and Vision: CLEF Beyond Text Retrieval Tasks



Gareth J. F. Jones

**Abstract** CLEF was initiated with intention of providing a catalyst to research in *Cross-Language Information Retrieval (CLIR)* and *Multilingual Information Retrieval (MIR)*. Focusing principally on European languages, it initially provided CLIR benchmark tasks to the research community within an annual cycle of task design, conduct and reporting. While the early focus was on textual data, the emergence of technologies to enable collection, archiving and content processing of multimedia content led to several initiatives which sought to address search for spoken and visual content. Similar to the interest in multilingual search for text, interest arose in working multilingually with multimedia content. To support research in these areas CLEF introduced a number of tasks in multilingual search for multimedia content. While investigation of image retrieval has formed the focus of the ImageCLEF task over many years, this chapter reviews tasks examining speech and video retrieval carried out within CLEF during its first 10 years, and overviews related work reported at other information retrieval benchmarks.

## 1 Introduction

Early work in *Cross-Language Information Retrieval (CLIR)* in the late 1990s focused on addressing the translation challenges of crossing the language barrier between formally authored text documents, such as news reports, and user search requests posed in various different languages. This growing interest in understanding and addressing the challenges of multilingual search coincided with the emergence of *Information Retrieval (IR)* research exploring search of archives of spoken and visual content. It soon became clear that these technologies could be used in combination to support CLIR for collections of spoken content (Sheridan et al. 1997; Jones 2001) and image content (Clough et al. 2006). This latter work

---

G. J. F. Jones (✉)

ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland  
e-mail: [Gareth.Jones@dcu.ie](mailto:Gareth.Jones@dcu.ie)

was an initial driver to the establishment of the ImageCLEF tasks at the *Conference and Labs of the Evaluation Forum (CLEF)* which provide a popular forum for the exploration of novel image retrieval tasks.

While investigation of CLIR for text collections revealed the language translation challenges which must be addressed for effective search, speech and video content was quickly shown to pose additional challenges. These arise from various sources and include the errors typically found in transcripts generated using *Automatic Speech Recognition (ASR)* technologies, difficulties of processing and recognising visual content, the temporal nature of audio-visual content which makes selection of accurate playback points a crucial issue for efficient access to relevant content, and the informal multi-topic structure of much of this content which makes definition of retrieval units uncertain with consequential impact on search effectiveness. An initial examination of this topic was presented in Jones (2000), which set out, among other ideas, an initial proposal for the study of crosslingual search of spoken content.

This chapter begins by considering the nature of audio-visual content and its indexing for search, and then provides summary descriptions of the three tasks at CLEF which focused on multimedia content and their findings. The first of these tasks looked at translation for a relatively straightforward *Cross-Language Spoken Document Retrieval (CL-SDR)* task for broadcast news, while the second was a more challenging speech retrieval task for interviews in the form of oral testimonies accompanied by rich sources of text metadata. The third task introduced studies examining crosslingual processing of audio-visual content. The chapter ends with a brief exploration of more recent work on these topics which have built on the findings of these CLEF tasks.

## **2 Searching for Speech and Video**

Search for speech and video archives presents challenges which go beyond those encountered in text IR. Perhaps the most obvious of these is that the contents of speech and video archives are not known; the words spoken, visual objects present and events contained or observed in them need to be labelled. Thus, the first stage of processing audio and visual content is to perform some form of content recognition, prior to being able to index the content into an IR system.

### **2.1 Indexing Audio and Visual Content**

Since they are different forms of media, the indexing of audio and visual content employs different feature extraction and representation technologies. This section gives a brief introduction to the established indexing methods for these data sources.



### 2.1.1 Audio Indexing

Perhaps the most obvious way of creating a transcript of spoken audio is to manually transcribe it. However, the cost of manual transcription coupled with the rapidly growing size of digital spoken archives has generally rendered this approach uneconomic. Thus, most work on spoken content search has made use of ASR systems to recognise words spoken related to search queries or form transcripts of the complete content.

ASR is a long standing area of investigation in the speech research community with a history dating back more than 40 years. While this period has seen consistent advances in the functionality and accuracy of these systems, their quality has seen significant improvement in recent years with the introduction of neural pattern recognition methods. Online web applications are making increasing use of speech interfaces, enabling the providers to collect vast amounts of spoken content to provide hitherto unimagined amounts of training data enabling improvements in ASR systems. Nevertheless, ASR systems still make transcription errors, in particular relating to unexpected words which fall outside their vocabulary, which cannot by definition be recognised correctly, and informal speech in uncontrolled or noisy environments. The rise of crowdsourcing platforms has enabled practical use of manual transcription at reasonable cost as an alternative to ASR. Mechanisms to do this and ensure transcription quality are described in Marge et al. (2010). Thus in some situations manual transcription can provide a practical alternative when ASR is not sufficiently accurate.

Whatever transcription method is used, the resulting transcript still contains errors which will impact on retrieval effectiveness. Most simply if a user enters a search word which has been spoken, but recognised incorrectly in the spoken content, retrieval accuracy will at best be different from that which would have been achieved with a perfect transcript since the matching score of these mismatched items will be reduced. The most common average effect is for the rank of relevant items to be reduced, although this effect is more complex at the level of individual documents (Eskevich and Jones 2014). One factor in this complexity is the often highly variable nature of speech where poor articulation, speaker cross-talk, and disfluencies can greatly degrade recognition accuracy of affected regions within individual recordings. These variations apply particularly in the case of unstructured spontaneous speech.

### 2.1.2 Video Indexing

Indexing of visual content for search presents multiple difficulties, the first being what should the content features of the index actually be, and how might these be identified? Searchers may wish to look for generic examples of visual features, e.g. people, buildings, objects, or named examples of them, or events occurring within a video, e.g. person running, an object falling. In the former case a visual recognition system would need to be able to recognise any feature to be searched

for, in the latter it would need to go further, not only identifying the features involved, but also interpreting how they are interacting within an identifiable event within a video. Indexing such features within a video has proven to be extremely challenging when seeking to do so beyond specific well defined events. For example, recognising the features and events within the clearly defined setting of a soccer match can be performed reliably, while processing of videos of ball-based sports in general cannot. Less specific features which could be used for diverse video sources have needed to focus on low-level features such as colours and textures, and easily defined shapes (Smeulders et al. 2000). The difference between how humans interpret, and generally want to search, visual content and what can be recognised by a visual recognition system is referred to as the *semantic gap*. A key ongoing challenge in visual indexing has been to reduce this semantic gap.

As with ASR, the application of neural pattern recognition methods in recent years has led to significant advances in visual processing greatly increasing the capabilities and accuracy of visual feature recognition. Crowdsourcing is also a potential method to improve the accuracy and semantic depth of visual indexing where humans provide labels to visual content (Rashtchian et al. 2010). However, it is difficult to anticipate what features users will include in visual queries, and combined with the huge range of interpretations of visual content, it is generally not possible to annotate visual content with regard to all potential interpretations.

## 2.2 *Fundamental Nature of Audio-Visual Content*

Aside from the need to index the audio-visual content, it is also important to take account of the nature of this content, in particular the fact that the content is *temporal*. This means that the natural way to interact with this content is to either listen to it, watch it, or some combination of these, generally with the content played back in real time. This is quite different to text and indeed static images, where the searcher is able to very rapidly browse very large amounts of content visually. Users of text retrieval systems are typically shown snippet summaries of text to gauge the relevance of retrieved content and be directed to potentially relevant material. Within large text items visual cues, e.g. by highlighting individual text features, can be used. In the case of images, humans are found to be able to scan very large numbers of thumbnail images very rapidly to identify items of interest.

The temporal nature of audio-visual content has important implications for content interaction in search applications. The speed of accessing the information in spoken content is constrained by the playback speed at which the listener comprehends the content, Speech can often be comprehended when played back at double its original speed, but listeners find this cognitively demanding and tiring, and may need to repeat details to properly assimilate them. This situation can be contrasted with text retrieval, where the speed with which content can be engaged with is limited only by the speed with which the searcher can read. The persistent nature of text also means that the searcher can scan the content repeatedly looking

for relevant material or checking individual details. A similar situation as that of engaging with spoken content arises in the case of video where users can visually scan speeded up video, but will find this demanding and may need to carefully review important details at real-time speed.

The time taken to review temporal content has led to significant research efforts into the exploration of user interface components for applications to support accessing audio-visual content. In the case of speech content these investigations have examined various forms of graphical representation of the content showing location of search words in the audio, or some or all of the speech transcript (Larson and Jones 2011). In the case of video search, the focus has been on storyboard representations, similar to 35 mm camera films of the content, containing keyframes extracted from the video to give searchers an overview of the content (Schoeffmann et al. 2010).

A step beyond these visual interfaces, which generally enable direct engagement with content and manipulation of interface widgets by users, is to attempt to identify where in the retrieved items the searcher would ideally start listening in order to access content relevant to their information need most efficiently, often referred to as *jump-in* points. Locating suitable points to start reviewing the content in an item requires both the identification of the relevant region within the audio-visual content and further, ideally, to determine where review should begin contextually to provide sufficient details for the content accessed to make sense to the user.

### 3 Evaluation of Audio-Visual Retrieval

In this section we begin with a brief overview of work in audio-visual retrieval prior to its introduction in tasks at CLEF. We then summarise the activities and outcomes of the three tasks at CLEF which have focused on spoken data and the use of the spoken information stream in video. Since it is of most relevance to CLEF, we focus here predominantly on work in speech retrieval. The section concludes with outlines of subsequent work in audio-visual retrieval which either builds directly on the CLEF tasks or is relevant to them.

#### 3.1 *Audio-Visual Retrieval Evaluation Prior to CLEF*

Interest in audio-visual retrieval began to emerge in the early 1990s enabled by advances in the storage of and online access to multimedia content, and analysis and recognition of spoken and visual content. We begin by examining early speech retrieval initiatives and then briefly consider video retrieval.

### 3.1.1 Beginnings of Speech Retrieval

Speech retrieval research began with the use of so called *keyword spotting* which was designed to recognise a small predefined set of words, typically around 50, which enabled topic based filtering of spoken documents. The first truly IR-based speech retrieval work was described in Glavitsch and Schäuble (1992), which proposed a subword-based indexing approach for German language content. Expansion in the vocabulary and recognition accuracy of ASR transcription systems around this time enabled the introduction of ASR-based transcripts in speech search (James 1995; Jones et al. 1996). A number of related studies were carried out at this time, most notably the Infromedia digital library system at Carnegie Mellon University (Hauptmann and Witbrock 1997).

The evaluation of retrieval effectiveness in all this work was based on private collections, and so no direct comparison of the indexing and retrieval methods used was possible. All of these studies developed tasks for which single topic documents could easily be identified for retrieval. For example, voice mail messages or news stories manually extracted from broadcast radio news. In this case, these form natural document-based retrieval units, and for this reason these studies are generally referred to as *Spoken Document Retrieval (SDR)*. While the collections used in these studies were generally very small, in the order of a few hours of data, it was clear that the vocabulary limitations and transcription errors inherent in ASR systems impacted on retrieval accuracy compared to accurate manual transcripts. However, researchers were encouraged by the consistent finding that SDR was relatively robust to such errors, an error rate of 30% in word recognition typically resulted in a reduction of retrieval *Mean Average Precision (MAP)* of around 10%.

To enable comparative evaluation of SDR methods and to expand the scale of the test collections used to evaluate them, in 1997 NIST introduced the TREC *Spoken Document Retrieval (SDR)* track. This ran for four cycles from TREC 6 to TREC 9 and developed from a known-item search task using around 50 h of broadcast news data to an ad hoc retrieval task using several hundred hours of broadcast TV and radio news. Results from the TREC 6 and TREC 7 SDR tasks confirmed earlier findings with regard to the relative effectiveness of manual and ASR transcripts, and in general showed word-based ASR transcripts to be the most effective speech indexing method. The document collections used for TREC 8 and TREC 9 did not have perfect manual transcripts, but relied on TV closed captions and rough manual transcripts of radio material as their reference transcript. Comparing retrieval effectiveness for the reference transcript with that for ASR transcripts for participant submissions showed that the application of various enhancement techniques in the IR process, including use of large contemporaneous news collections for term weighting and query expansion, found similar MAP values for both transcript types. Thus, essentially when using appropriate IR techniques manual transcripts and ASR-based transcripts, SDR was found to have achieved similar retrieval effectiveness. Although it is worth noting that while the best ASR transcripts had a state-of-the-art error rate for that time of a little more than 20%, the imperfect reference transcripts based on TV closed captions had a measured error

rate of 14.5% and the rough manual transcripts of the radio broadcasts a measured error rate of 7.5%. The organisers of TREC SDR declared SDR essentially a solved problem, and research interest for multimedia retrieval benchmarks moved to the task of video retrieval. A detailed review of the TREC SDR track is contained in Garofolo et al. (2000).

A comprehensive introduction to speech retrieval with an extensive list of references can be found in Larson and Jones (2011).

### 3.1.2 The Emergence of Video Retrieval

Similar to SDR, work in Video Retrieval began with independent research efforts in the 1990s. This work examined topics including:

- simple visual features based on colours and textures, detectors for features such as faces, optical character recognition (OCR) within video frames;
- temporal analysis, looking at individual frames, identifying shot beginning and ending points associated with changes of camera, locating scenes where a sequence of shots has semantic coherence, use of complete videos;
- interfaces for video browsing; and
- querying using text queries, image queries and combinations of these.

As the investigation of text querying suggests, much of this work was very dependent on the availability of a transcript of the spoken information stream of the video. Researchers were of course interested in search of video which did not have a meaningful associated spoken soundtrack, where the emphasis would need to be on the visual information stream, but this proved very challenging and was not generally found to be effective at that time. A good overview of this work is contained in Marchand-Maillet (2000), with a detailed review of visual information processing at this time in Smeulders et al. (2000).

Again similar to SDR, with the desire to enable comparative evaluation and to produce research resources, often too expensive for individual research groups to create, NIST introduced the *TREC Video Retrieval Evaluation (TRECVID)* track in 2001. This was soon expanded and moved to a separate campaign with its own workshop separate from the main TREC campaigns which focused on search tasks for text-based content.<sup>1</sup> The early years of TRECVID, prior to the introduction of a video search task at CLEF 2008, involved consolidation of the earlier independent activities. Thus, there were tasks examining automatic shot boundary detection, identification of named features including objects and people, and primarily the construction and evaluation of interactive video search system. Participants were provided with multimodal search topics combining text and images, and required to use their interactive systems to locate relevant items. It should be noted that the topics used in these investigations focused primarily on satisfaction of visual

---

<sup>1</sup>[trecvid.nist.gov](http://trecvid.nist.gov).

information needs, rather being truly multimodal in nature. In these evaluations the shot was used as the retrieval unit, while these were easy units to define, it is not clear how useful retrieval of these items is to users who might prefer more semantically defined units or recommended playback points. Participants could typically be required to locate as many items relevant to a given search topic as possible within a fixed time period. The interactive systems enabled exploration of the use of visual features in conjunction with ASR transcripts and standard IR methods enhanced with techniques such as visual and text based relevance feedback. They also enabled investigation of user interfaces for video retrieval. A detailed summary overview of the first 6 years of the TRECVID campaigns is contained in Smeaton et al. (2006).

## **3.2 Audio-Visual Retrieval @CLEF**

This section provides summary descriptions of the audio-visual IR tasks explored at CLEF. These tasks were carried out in three phases between CLEF 2002 and CLEF 2009: CLEF 2002–2004: *Cross-Language Spoken Document Retrieval (CL-SDR)*, CLEF 2005–2007 *Cross-Language Speech Retrieval (CL-SR)*, and CLEF 2008–2009 VideoCLEF.

### **3.2.1 CLEF 2002–2004: Cross-Language Spoken Document Retrieval (CL-SDR)**

Following the success of the cross-language text retrieval tasks at the first two CLEF workshops in 2000 and 2001, it was decided to begin an exploration of CL-SDR at CLEF 2002. The initial CL-SDR task opted for a low entry cost activity of using the TREC SDR test collections from TREC 8 and TREC 9. These monolingual test collections were extended for cross language investigation by using native speakers to translate the search topics into French, German, Italian, Spanish and Dutch.

For CLEF 2002 and CLEF 2003, the CL-SDR task focused on the document retrieval task from the TREC SDR tasks where the transcripts had been manually segmented into news stories which formed the target retrieval documents. Retrieval effectiveness for this task was simply evaluated using MAP. The primary comparison was between retrieval performance for the original monolingual SDR task and that for the CL-SDR task using the translated queries. Translation for CL-SDR was examined by participants using various methods. While it was found that using multiple translation resources in combination produced more robust retrieval, MAP was typically reduced by on the order of more than 20% for CL-SDR compared to the participant's monolingual baseline. Similar to the earlier work in the original TREC SDR task, use of external text resources was found to be effective in improving retrieval effectiveness, although it was also demonstrated that, at least in the domain of broadcast news, this text material must be temporally matched to

the target retrieval collection in order to produce a beneficial effect on retrieval. The details of these investigations are described in Jones and Federico (2003), Federico and Jones (2004).

For CLEF 2004, the CL-SDR task focused on the more demanding unknown story boundary condition from TREC SDR. In this setting, the transcripts are provided without story boundaries, and participants must automatically divide the transcripts to form units for retrieval as part of their SDR system. Retrieved units are then judged relevant if they overlap with relevant manually segmented content. In this condition MAP between monolingual SDR and CL-SDR was found to be degraded on the order of 40% when no external text data was used to support retrieval, although this difference was reduced for all languages when external text was used, where it should be noted that absolute monolingual SDR effectiveness was also improved by the use of external text data. More detail of the CLEF 2004 CL-SDR task is available in Federico et al. (2005).

While the CL-SDR tasks at CLEF 2002–2004 showed significant reduction in MAP arising from translation problems, which were to some extent addressed by the use of multiple translation resources, the task was stopped without any serious investigation into solving this problem.

### 3.2.2 CLEF 2005–2007: Cross-Language Speech Retrieval (CL-SR)

While the first CL-SDR task identified significant problems in translation, the underlying broadcast news retrieval task is actually a relatively easy speech search task. This is due to a number of reasons:

- The content can be segmented into natural document units for retrieval.
- There is significant training data available to develop effective ASR systems adapted to the target dataset.
- The content can be considered to be “self describing” in the sense that each news article has to provide details of the main actors and their roles, locations, events being described, etc., so that the listener knows the background and details of the story. These characteristics often make these news story segments likely to match queries seeking them very well, making them easy to be reliably retrieved.
- Since multiple broadcast news stories were used and many stories develop slowly over a number of news broadcasts, there are generally multiple relevant documents for each query. Thus even if not all relevant documents are retrieved at high rank, MAP will often only be slightly reduced relative to a manual transcript.
- There are large amounts of closely related textual documents available to better estimate parameters of retrieval systems, and apply techniques such as query expansion.

Partially in recognition of this, CLEF 2005 saw the introduction of a much more challenging task of searching a collection of oral testimonies which ran for three editions from CLEF 2005–2007. These testimonies were interviews with Holocaust survivors, witnesses and rescuers collected by the Survivors of the Shoah

Visual History Foundation. Since the oral testimonies were extended recordings of interviews sometimes lasting 2 h, there were no naturally occurring documents. Speech retrieval in the absence of clearly definable document units can be referred to as *Spoken Content Retrieval (SCR)*, reflecting the more general task of searching for relevant spoken content. The label SCR not being in general usage at the time, this new task at CLEF was named CL-SR. Automatic transcripts of the content were produced as part of the Multilingual Access to Large Spoken Archives (MALACH) project (Byrne et al. 2004). The absence of a manual transcript meant that it was not possible to compare manual and ASR retrieval effectiveness in these tasks.

The data collection for the CL-SR task at CLEF 2005 was in English and consisted of 272 interviews totalling 589 h of speech data. The interviews were manually divided into 8104 semantically coherent segments. It should be noted that this segmentation was performed carefully by subject matter experts. Thus, such segments would not normally be available for an archive of this type, and the expectation was that participants should be able to search for relevant content in the unsegmented audio files. Participants were provided with multiple sources of information for each manually labelled segment of an interview as follows:

- INTERVIEWDATA field: this gave details of the interviewee, and was the same for all segments in an interview.
- NAME field: containing the names of other persons mentioned in the segment.
- ASRTEXT field: containing two ASR transcripts of the interview. One created using an ASR system tuned for recognition of this dataset with an average word error of 38% and named entity error rate of 32% and the other created using an earlier ASR system with an average word error rate of 40% and named entity error rate of 66%.
- MANUALKEYWORDS field: containing thesaurus descriptors manually assigned by subject domain experts from a large domain specific thesaurus. An average of about five descriptors were assigned to each segment.
- SUMMARY field: containing a three sentence summary created by a subject matter expert using free text addressing the questions: who? what? when? where?
- Two AUTOKEYWORD fields: containing thesaurus descriptors automatically assigned using two variants of a k-Nearest Neighbour classifier.

Elaborate procedures were used to create the search topic sets and corresponding relevance information for the training and evaluation phases. These procedures are described in detail in the task overview paper (White et al. 2006). In summary, 38 training and 25 evaluate topics were created. All topics were originally written in English, and were re-expressed in Czech, French, German and Spanish. Rather than the standard pooling method used to create IR test collections, relevance assessment was carried out by subject experts using a search-guided relevance judgment procedure.

Participants were required to submit a monolingual run created using only automatically derived index information. They were further allowed to submit runs using any combination of fields they wished, and were encouraged to investigate the use of non-English topics. A key finding of the participant results was that the



best run using manual metadata yielded a statistically significant improvement over the strongest results obtained using only automatically created data, including the ASR transcripts. Cross language search generally resulted in significant reduction in MAP of between 20 and 50%, although curiously there were examples of cross language search producing small improvements in MAP.

The CL-SR task at CLEF 2006 maintained the same document collection as the CLEF 2005 task, but introduced 42 new topics, 33 of which were ultimately used in the official evaluation. These were again translated into Czech, Dutch, French, German and Spanish. Relevance assessment was carried out in the same manner as 2005, and again performed by subject experts. Evaluated results using automated indexing methods were lower in absolute terms for the 2006 topic set as compared to the 2005 results. This appeared to suggest that the topics were less well matched to the ASR transcripts than in 2005, since results for manual data indicated that the topics were not generally harder. Cross language results were on the order of 20% below those for corresponding monolingual runs.

A significant change for the CL-SR task at CLEF 2006 was the introduction of a second collection of interviews in Czech. Aside from the change of language, the key difference was that no manual segmentation was performed on the Czech transcripts. This led to a search task that was time-orientated rather than segment-orientated. In this task participants were asked to identify replay start times for relevant content, rather than to identify relevant predefined segments. Results were returned in form of ranked lists of playback start times (or jump-in points), rather than document identifiers.

In the absence of manual segmentation into semantic regions, the transcripts were automatically segmented into passages of 3 min duration, with start times spaced by 1 min. Running a script created to do this, 11,377 overlapping passages were identified.

The Czech information provided to participants was as follows:

- **DOCNO** field: a unique document number in the same format as the playback start times to be returned by the IR system.
- **ASRSYSTEM** field: specifying which of two available ASR transcripts were to be used; referred to as “2004” and “2006”, where 2006 was a later and better system.
- **ASRTEXT** field: containing the selected ASR transcript from the passage beginning specified in **DOCNO**.
- **ENGLISHAUTOKEYWORD** field: 20 automatically selected thesaurus terms for the passage.
- **CZECHAUTOKEYWORD** field: Czech translation of the **ENGLISHAUTOKEYWORD** field.
- **INTERVIEWDATA** field: name of the person being interviewed, the same for all passages of the same interview.
- **ENGLISHMANUALKEYWORD** field: manually assigned thesaurus terms assigned to each passage by subject domain experts.

- **CZECHMANUALKEYWORD** field: Czech translation of the English thesaurus terms in the **ENGLISHMANUALKEYWORD** field.

The topic set was the same 115 topics provided for the English language task manually translated into Czech by native speakers, with the addition of 10 specially adapted topic statements.

Results were evaluated using a variant of mean Generalised Average Precision (mGAP) (Kekäläinen and Järvelin 2002). This metric was essentially a combination of MAP multiplied with a linear decaying function whose value depended on the distance in either direction, late or early, between the system-recommended start time for the content in the participant's submission and that of relevant content identified by a manual assessor. An error between these values of more 150s was treated as a no-match condition.

Relevance assessments were created for 29 Czech topics using a combination of search-guided relevance assessment and a pool of highly ranked start points provided by participants. A total of 1322 start times of relevant passages was identified, yielding an average of 46 relevant passages per topic.

The complexity of the data management and task design meant that results in the first year of this new Czech task were generally inconclusive, although it was found that the automatically assigned keywords were not helpful in retrieval. Full details of the CLEF 2006 CL-SR task are available in Oard et al. (2007).

For the final edition of the CL-SR task at CLEF 2007, the same English and Czech tasks were examined using the same document sets with largely the same topic statements and relevance assessments. In the case of the English task, exactly the same test collection was used with participants instructed not to develop their systems using the test topics. For the Czech task, the 29 topics from CLEF 2006 were used for training, with a new set of 42 test topics with relevance assessment carried out using the same protocol as used for this task at CLEF 2006. Full details of the CLEF 2007 CL-SR task are available in Pecina et al. (2008).

### 3.2.3 CLEF 2008–2009: VideoCLEF

Following the conclusion of the CL-SR task at CLEF 2007, it was decided to develop a task focusing on cross language search involving video. VideoCLEF was design to extend the achievements of the CL-SR task to incorporate the challenges of video search, and to be complementary to TRECVID. While TRECVID at this time focused on what was depicted in a video, the goal of VideoCLEF was to develop and evaluate tasks involving the analysis of multilingual video content. In particular it worked with *dual language* video in which two languages are spoken, but the languages do not duplicate the content. Examples of dual language video include documentaries where interviewees do not speak the dominant language of the show, but rather speak another language.

The video data used for VideoCLEF at CLEF 2008 was supplied by the *Netherlands Institute of Sound and Vision (NISV)*,<sup>2</sup> one of the largest audio/video archives in Europe. The video used was dual language content predominantly in Dutch featuring English speaking experts and studio guests.

The VideoCLEF task at CLEF 2008 was referred to as *Vid2RSS*. The main subtask was a classification task which required participants to automatically assign thematic high-level semantic features in the form of subject category labels to this dual language video. The labels used for this task were a subset of those used by archive staff at the NISV for annotation of archival materials. The use of these labels was chosen since manually assigned subject labels were available for this data providing a gold standard for evaluation in the task.

In addition to the classification subtask, there were two other subtasks: a translation subtask and a keyframe extraction subtask. The translation subtask required participants to translate topic-based feeds from Dutch into a target language. The feeds consisted of a concatenation of the video's title, a short description derived from archival metadata and a keyframe representing the video's content. Evaluation was carried out using human assessment of adequacy and fluency, where all assessors had high-level mastery of the source and target languages. One of the main problems for this task was the frequent failure to translate Dutch compound words. The keyframe extraction subtask required participants to select a keyframe which best represents the semantic content of the video from among a provided set.

The CLEF 2008 VideoCLEF task successfully demonstrated that classification of dual language TV documentaries into subject classes was a challenging and interesting task. It was found to be a more difficult task than similar classification of broadcast news, this was believed to be due to the unscripted nature of much of the speech in interviews and discussions. A more detailed description of the tasks and findings of VideoCLEF 2008 can be found in Larson et al. (2009).

VideoCLEF at CLEF 2009 expanded on the exploratory tasks introduced in 2008 and offered three tasks. These used two datasets both containing Dutch-language television programmes, predominantly documentaries with inclusion of talk shows. Much of the material was thus of an informal unscripted conversational nature making it more challenging than earlier work using broadcast news data outlined above. The Classification Task made use of data provided by the NISV which had previously been used at TRECVID in 2007 and 2008. The Affect Task and the Linking Task used a second dataset also supplied by the NISV for the documentary series *Beeldenstorm*. This consisted of 45 short-form Dutch documentaries lasting about 8 min each on subjects in the visual arts. The following three subtasks were investigated:

*Subject Classification Task:* This required participants to automatically tag videos with subject theme labels, e.g. 'physics', 'culture', 'poverty'. The purpose of assigning these labels was to make the videos more findable to users. The task had the specific goal of reproducing the subject labels hand assigned to the videos by

---

<sup>2</sup>[www.beeldengeluid.nl](http://www.beeldengeluid.nl).

archivists at the NISV. The Subject Classification in 2009 represented a significant expansion of the exploratory task introduced in 2008. The number of videos was increased from 50 to 418, and the number of labels to be assigned expanded from 10 to 46. The effectiveness of the assignment was measured using MAP.

*Affect Task*, also referred to a ‘narrative peak detection’ task: This involved automatically detecting dramatic tension in short form documentaries. Narrative peaks were defined to be those places in a video where viewers report feeling a heightened emotional effect due to dramatic tension. The Affect Task was intended as a first step towards investigation of video content with respect to characteristics important to viewers, but not related to the video topic. Narrative Peaks should be differentiated from cases such as “hotspots” in videos where participants depicted in the video are highly engaged with their situation, and can actually self report being so; narrative peaks relate to the reaction of the viewer to the video being observed.

*Linking Task*: This required participants to automatically link video to semantically related Web content in a different language. The task involved linking episodes of the Dutch language *Beeldenstorm* documentary to English language Wikipedia articles about subjects related to the video. Participants were supplied with a set of 165 multimedia anchors, short video segments of about 10 s duration, located in the documentaries. For each anchor, participants had to return a list of Wikipedia pages ranked by potential relevance. It should be noted that this went beyond a named-entity linking task, an anchor may include mention of a named-entity or it may not, the task was more akin to a standard IR task with the anchor as a query and the topical information potentially split between the visual and speech channels. The success of link creation was evaluated using Mean Reciprocal Rank (MRR).

VideoCLEF in 2009 successfully investigated the challenges of the three tasks with participants in the Subject Classification Task and the Linking Task focusing on information in the speech channel, and participants in the Affect Task giving greater emphasis to the exploitation of combinations of the audio, spoken and visual channels (Larson et al. 2010).

### ***3.3 Audio-Visual Search Evaluation Post CLEF***

Interest in audio-visual search did not end with the conclusion of the VideoCLEF task in 2009. Many of the findings of the speech and video tasks at CLEF served mainly to identify the challenges of developing technologies to support effective audio-visual search and its evaluation, raise detailed research questions which needed to be addressed, and highlight the shortcomings of the technologies available at that time. Recent years have seen significant advances in both ASR and video indexing technologies meaning that more accurate and richer indexing of multimedia content has become available, with consequential improvements both in the effectiveness of established search tasks, but, perhaps more excitingly, the potential for a diverse and extensive range of new possible applications.

TRECVID has continued to innovate and explore tasks in video search incorporating the examination of the impact of improved video indexing methods to enhance the effectiveness of existing tasks and establish new ones. This section first briefly introduces the *MediaEval*<sup>3</sup> multimedia benchmark initiative which commenced in 2010 following the conclusion of audio-visual search tasks at CLEF. MediaEval is designed to provide a forum for the exploration and evaluation of new and emerging task possibilities. We then briefly review tasks which have appeared within existing benchmarks or new evaluation initiatives which directly extend work examined within the audio-visual search tasks at CLEF.

### 3.3.1 MediaEval

The *MediaEval* multimedia evaluation benchmark runs on an annual cycle similar to CLEF, celebrating its 10th anniversary in 2019. MediaEval typically runs about eight tasks in each edition selected following an open call for proposals. Tasks in the first 10 years of MediaEval focused on images, video, speech, and music, covering topics including use of context information such as location, exploration of issues of affect in multimedia, querying classical music scores and automatically creating chronologically-ordered outlines of events captured in multimedia social media archives.

Of particular relevance to the speech retrieval tasks carried out at CLEF, are the Rich Speech Retrieval (RSR) (Larson et al. 2011) and Search and Hyperlinking (S&H) (Eskevich et al. 2012a, 2013, 2014, 2015) tasks which extended investigations of search of unstructured spoken content carried out at CLEF.

The RSR task took up the challenges introduced in the CLEF 2005–2007 CL-SR task to explore search of diverse content of varying levels of formality and preplanning. The `blip.tv` collection used for this task is an archive of semi-professional user generated content (UGC) in the form of video crawled from the internet (Schmiedeke et al. 2013). As UGC content, there is no control over content design, content semantics, structure or form of the content, recording conditions, articulation quality or clarity of the speakers and no scripts or other documentation available. Participants in the RSR task were thus required to develop search solutions which were robust to these characteristics of the data. Participants were provided with a noisy ASR transcript of the content and a set of search queries. The task was to identify and retrieve relevant content from within the audio recordings by identifying jump-in points at which to begin playback. The development of both the RSR and S&H tasks involved extensive innovative use of crowdsourcing methods for the creation of search queries and labelling of relevant content (Jones 2013). The evaluation metric was a variation of mGAP based on MRR rather than MAP, thus only the first relevant jump-in point found in the retrieved list of proposed starting times was considered. The main challenge taken up in the task submissions

---

<sup>3</sup>[multimediaeval.org](http://multimediaeval.org).

was to explore alternative methods of automated identification of retrieval units, the principal contrast being between methods which took the simple approach of extracting overlapping fixed length segments and topical segmentation methods which sought to identify semantically coherent segments as the retrieval units. The perhaps surprising conclusion being that the simple fixed length segmentation method was more effective. More detailed subsequent analysis of this finding provided a proper understanding of these results (Eskevich et al. 2012b).

The S&H task expanded on the scope of the RSR to include both speech retrieval, and an innovative new subtask exploring automated video-to-video linking. The long term vision of this latter task was the creation of indexing systems which can automatically identify areas of “interest” within a video and use these as anchor points to form search queries linking to “related” video content within the archive. A user engaging with such a video would then be able to follow these links to related video material. From one perspective, this might be considered a more advanced version of the Linking Task from VideoCLEF at CLEF 2009. While an appealing and perhaps intuitively simple idea, much of the work in the development of the Hyperlinking subtask focused on exploring the definition and role of these links (Aly et al. 2013; Ordelman et al. 2015). The S&H task worked first with the blip dataset used for the RSR task (Eskevich et al. 2012a), and then using an archive of broadcast TV provided by the BBC where a full ad hoc search task was established (Eskevich et al. 2013, 2014, 2015). The first three editions of the S&H task used manually specified anchor points with the videos as the starting point of the video links, while the final edition at MediaEval introduced a new subtask challenge in which participants had to take a step towards a fully automated video hyperlinking system by seeking to automatically identify desirable anchor points to act as the starting points for video links (Eskevich et al. 2015). After four successful editions running within MediaEval, the Video Hyperlinking element of the S&H task ran for three editions as a separate task within TRECVID (Over et al. 2015; Awad et al. 2016, 2017).

The MediaEval RSR and S&H tasks were entirely monolingual and contained only English language content. The data from the S&H 2012 Search task was later extended for exploration of CL-SR. This was done by first adapting the task to an ad hoc search task. This was achieved by first modifying the queries to be less specific to increase the scope of the relevant documents, then performing searches with these modified queries using multiple different strategies, and finally carrying out pooled relevance assessment of the retrieved documents. Separately these revised queries were manually translated into French and Arabic by native speakers (Khwileh and Jones 2016; Khwileh et al. 2017). This is to date the only study of CL-SR for spoken UGC. Results for these studies revealed the expected issues in terms of errors in ASR transcripts of the audio, and problems of translation for CLIR, in particular in this case for suitable translation of Arabic named entities. The greatest challenge identified in this work however probably relates to the length of the content. Since this is UGC, there is no managed editorial control of the length of the items, and the transcripts were found to vary in length from a few tens of words to more than 20,000, with an average length of 700 words. This wide variation in

document length poses well established challenges for IR in terms of the reliable placement of relevant items at high rank. This issue is particularly important for multi-topic spoken content where the automated identification of jump-in points to support access to relevant content efficiently is highly desirable. This requires the segmentation of the data to locate relevant content and jump-in with high precision. However, the lack of consistency in the format of the data means that application of a single segmentation approach for all content items is found to produce suboptimal results (Khwileh and Jones 2016). Further work on the effective retrieval of this type of content is required to address the issues identified in the work carried out to date.

### 3.3.2 NTCIR SpokenDoc and Spoken Query&Doc tasks

The subject of SCR was also taken up in a series of tasks at *NII Testbeds and Community for Information access Research (NTCIR)* from 2010 to 2016. The SpokenDoc task at NTCIR-9 and NTCIR-10 offered a Japanese language SCR task focused on retrieval of relevant content from within recordings of technical lectures and workshop presentations (Akiba et al. 2011, 2013). The Spoken Query&Doc task at NTCIR-11 and NTCIR-12 extended this to incorporate search using spoken queries (Akiba et al. 2014, 2016). This latter task remains the only one to explore the use of spoken queries in an SCR benchmark task, in addition to the textual queries used in all previous SCR tasks.

The SCR tasks at NTCIR were particularly notable since accurate manual transcripts were available for all the content to be searched. This enabled reliable analysis of the impact of ASR errors on retrieval behaviour. In addition, the content was divided into segments referred to as “inter-pausal units” (IPUs) which were created by dividing the speech at points where there is a pause no shorter than 200ms. IPUs are though too small to serve as retrieval units, and participants explored various approaches to the creation of segment units for retrieval, similar to the earlier studies described above, these again focused on overlapping fixed length units and lexical segmentation methods. A key feature of the NTCIR tasks was that relevance assessment was carried out at the level of IPUs, this meant that it was possible to carry out detailed examination of the retrieval of relevant content. For example, to look at the length of a retrieval unit extracted from the transcript, its overlap (if any) with neighbouring units, and the proportion of content contained in each unit labeled as relevant. Evaluation was based on variants of MAP taking into account the relevant and non-relevant IPUs in retrieved passages.

The key finding in these tasks was once again that simple fixed length overlapping content units were more effective for retrieval than more complex units extracted using lexical segmentation methods. In this case, the detailed annotation of the content meant that it was possible to carry out careful analysis of the reasons underlying the results. However, while comparison and analysis of the results improved understanding of the behaviour of SCR methods, none of the systems proposed appears to represent the best potential solution for all situations with all effective methods having strengths and weaknesses for retrieval with individual queries.

The NTCIR tasks created valuable research resources. For example, these have enabled the detailed investigation of the relationships between word error rate in individual retrieved segments (which was observed to vary significantly between segments), and retrieval behaviour. A noticeable correlation was found between word error rate and document rank in Sanderson and Shou (2007), where it was observed that documents with lower word error rate tended to be ranked higher by the IR system, regardless of document relevance. Further, while work described in Eskevich and Jones (2014) confirms the standard result that ASR errors lead to a reduction in MAP as observed in many other studies, it also noted that the actual impact on the user experience in terms of the promotion and demotion of items in a ranked retrieval list between accurate and ASR transcripts arising from transcription errors was dramatic and unpredictable.

Contextualisation techniques seek to improve the rank of a relevant element by considering information from its surrounding elements and its container document. The NTCIR Spoken Query&Doc test collections have been used to study the application of alternative contextualisation methods for SCR (Racca and Jones 2016). In addition to the general utility of contextualisation techniques in SCR, this study explored their potential to provide robustness to ASR errors in segments by taking account of the contents of adjacent segments. The reported experiments demonstrate that context information becomes increasingly valuable for improving retrieval effectiveness as ASR errors increase. Once again though, variations in individual situations mean that a simple fixed parameter implementation of this method appears not to realise its full potential.

### 3.3.3 NIST OpenCLIR Evaluation

As outlined above, the last 20 years has seen significant advances in the understanding of SCR and improvements in its component technologies. Most of this work has been carried out using the best available systems, which in the area of language technologies, means those with the greatest investment in analysis of the languages involved and development of application training resources. These languages are typically those spoken by large populations in technologically advanced countries, there are though a very large number of languages spoken either by smaller communities or where there is currently little economic interest in them. Such languages, referred to as *low-resource* languages can unpredictably and very rapidly become of great interest, for example in the event of a natural disaster.

It is highly desirable in such situations for those engaged in supporting local communities to be able to access information in local languages and to render them in more widely spoken languages. Mindful of this desire, the goal of the OpenCLIR (Open Cross Language Information Retrieval) evaluation<sup>4</sup> organised by NIST in 2019 was to develop methods to locate text and speech content in “documents”

---

<sup>4</sup>[www.nist.gov/itl/iad/mig/openclir-evaluation](http://www.nist.gov/itl/iad/mig/openclir-evaluation).



(speech or text) in low-resource languages using English queries. The OpenCLIR evaluation was created out of the larger and more ambitious IARPA MATERIAL<sup>5</sup> program. The purpose of OpenCLIR was to provide a simplified, smaller scale evaluation open to all researchers. Participants in the OpenCLIR evaluation were provided with some limited information of the linguistic characteristics of the target language, examples of bitexts (sentences in the language and corresponding English translations), and manual transcripts of audio and files of audio content. Using these resources and potentially others available to them from elsewhere, participants had to build CL-SR systems. The amounts of training materials provided are much less than those used to construct state-of-the-art tools for well resourced languages, and developed systems thus suffer from extensive translation and transcripts problems. The objective of participants then was to address these challenges within the resources available. Many of these are actually those faced in early work on CLIR and SDR where resources were limited for all languages, and it is interesting to consider how the creative solutions proposed in this early work might find utility in settings such as the OpenCLIR task where the latest technologies cannot be applied.

## 4 Concluding Remarks

At the time of the first edition of CLEF in 2000, SDR had been declared a largely solved problem and video indexing was emerging as a major area of research. The tasks carried out within CLEF and related subsequent and parallel activities at MediaEval and TRECVID respectively demonstrated that SDR, or more ambitiously the broader scoped topic of SCR, represented a much greater challenge than previously believed and was far from being a solved problem, and that the potential for accessing and exploiting audio-visual content goes far beyond developing text IR systems adapted for non-text media. Work to date is beginning to offer operational solutions for some of the more straightforward potential applications, but much remains to be done to realise the potential of this data to support and enrich the life activities of users in areas as diverse as entertainment, education, health and humanitarian efforts.

**Acknowledgements** The success of the CLEF and MediaEval tasks described in this chapter would not have been possible without the work of the task co-chairs Marcello Federico, Douglas W. Oard, Martha Larson, Maria Eskevich, Robin Aly and Roeland Ordelman.

---

<sup>5</sup>[www.iarpa.gov/index.php/research-programs/material](http://www.iarpa.gov/index.php/research-programs/material).

## References

- Akiba T, Nishizaki H, Aikawa K, Kawahara T, Matsui T (2011) Overview of the IR for spoken documents task in NTCIR-9 workshop. In: Kando N, Ishikawa D, Sugimoto M (eds) Proceedings of the 9th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access. National Institute of Informatics, Tokyo
- Akiba T, Nishizaki H, Aikawa K, Hu X, Itoh Y, Kawahara T, Nakagawa S, Nanjo H, Yamashita Y (2013) Overview of the NTCIR-10 spokendoc-2 task. In: Kando N, Kishida K (eds) Proceedings of the 10th NTCIR conference on evaluation of information access technologies. National Institute of Informatics, Tokyo
- Akiba T, Nishizaki H, Nanjo H, Jones GJF (2014) Overview of the NTCIR-11 spokenquery&doc task. In: Kando N, Joho H, Kishida K (eds) Proceedings of the 11th NTCIR conference on evaluation of information access technologies. National Institute of Informatics, Tokyo
- Akiba T, Nishizaki H, Nanjo H, Jones GJF (2016) Overview of the ntcir-12 spokenquery&doc-2 task. In: Kando N, Sakai T, Sanderson M (eds) Proceedings of the 12th NTCIR conference on evaluation of information access technologies. National Institute of Informatics, Tokyo
- Aly R, Ordelman R, Eskevich M, Jones GJF, Chen S (2013) Linking inside a video collection - what and how to measure? In: Proceedings of the first worldwide web workshop on linked media (LiME-2013), International World Wide Web Conference Committee (IW3C2), Geneva
- Awad G, Fiscus J, Joy D, Michel M, Smeaton AF, Kraaij W, Eskevich M, Aly R, Ordelman R, Jones GJF, Huet B, Larson M (2016) TRECVID 2016: evaluating video search, video event detection, localization, and hyperlinking. In: The sixteenth international workshop on video retrieval evaluation (TRECVID 2016). National Institute of Standards and Technology (NIST), Special Publication 500-321, Washington
- Awad G, Butt A, Fiscus J, Joy D, Delgado A, McClinton W, Michel M, Smeaton A, Graham Y, Kraaij W, Quénot G, Eskevich M, Roeland Ordelman GJFJ, Huet B (2017) Trecvid 2017: evaluating ad-hoc and instance video search, events detection, video captioning, and hyperlinking. In: The seventeenth international workshop on video retrieval evaluation (TRECVID 2017). National Institute of Standards and Technology (NIST), Special Publication 500-321, Washington
- Byrne W, Doermann D, Franz M, Member S, Gustman S, Soergel D, Ward T, jing Zhu W (2004) Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Trans Speech Audio Process* 12(4):420-435
- Clough P, Sanderson M, Reid N (2006) The Eurovision St Andrews collection of photographs. *SIGIR Forum* 40(1):21-30
- Eskevich M, Jones GJF (2014) Exploring speech retrieval from meetings using the AMI corpus. *Comput Speech Lang (Special Issue on Information Extraction and Retrieval)* 28(5):1021-1044
- Eskevich M, Jones GJF, Chen S, Aly R, Ordelman R, Larson M (2012a) Search and hyperlinking task at mediaeval 2012. In: Larson MA, Schmiedeke S, Kelm P, Rae A, Mezaris V, Piatrik T, Soleymani M, Metz F, Jones GJF (eds) Working Notes Proceedings of the MediaEval 2012 multimedia benchmark workshop. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, <http://ceur-ws.org/Vol-927/>
- Eskevich M, Jones GJF, Larson M, Wartena C, Aly R, Verschoor T, Ordelman R (2012b) Comparing retrieval effectiveness for alternative content segmentation methods for internet video. In: Proceedings of the 10th workshop on content-based multimedia indexing. IEEE, New Jersey, CBMI 2012
- Eskevich M, Jones GJF, Chen S, Aly R, Ordelman R (2013) The search and hyperlinking task at mediaeval 2013. In: Larson M, Anguera X, Reuter T, Jones GJF, Ionescu B, Schedl M, Piatrik T, Hauff C, Soleymani M (eds) Working notes proceedings of the MediaEval 2013 multimedia benchmark workshop. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, <http://ceur-ws.org/Vol-1043/>

- Eskevich M, Aly R, Racca DN, Ordelman R, Chen S, Jones GJF (2014) The search and hyperlinking task at mediaeval 2014. In: Larson M, Ionescu B, Anguera X, Eskevich M, Korshunov P, Schedl M, Soleymani M, Petkos P, Sutcliffe R, Choi J, Jones GJF (eds) Working notes proceedings of the MediaEval 2014 multimedia benchmark workshop. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, <http://ceur-ws.org/Vol-1263/>
- Eskevich M, Aly R, Ordelman R, Racca DN, Chen S, Jones GJF (2015) SAVA at Mediaeval 2015: Search and anchoring in video archives. In: Larson M, Ionescu B, Sjöberg M, Anguera X, Poignant J, Riegler M, Eskevich M, Hauff C, Sutcliffe R, Jones GJF, Yang YH, Soleymani M, Papadopoulos S (eds) Working notes proceedings of the MediaEval 2015 multimedia benchmark workshop. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, <http://ceur-ws.org/Vol-1436/>
- Federico M, Jones GJF (2004) The CLEF 2003 cross-language spoken document retrieval track. In: Peters C, Braschler M, Gonzalo J, Kluck M (eds) Comparative evaluation of multilingual information access systems: fourth workshop of the cross-language evaluation forum (CLEF 2003) revised selected papers. Lecture notes in computer science (LNCS), vol 3237. Springer, Heidelberg, p 646
- Federico M, Bertoldi N, Levow GA, Jones GJF (2005) CLEF 2004 cross-language spoken document retrieval track. In: Peters C, Clough P, Gonzalo J, Jones GJF, Kluck M, Magnini B (eds) Multilingual information access for text, speech and images: fifth workshop of the cross-language evaluation forum (CLEF 2004) revised selected papers. Lecture notes in computer science (LNCS), vol 3491. Springer, Heidelberg, pp 816–820
- Garofolo JS, Auzanne CGP, Voorhees EM (2000) The trec spoken document retrieval track: a success story. In: Content-Based Multimedia Information Access - vol 1, LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, Paris, France, France, RIAO '00, pp 1–20
- Glavitsch U, Schäuble P (1992) A system for retrieving speech documents. In: Belkin NJ, Ingwersen P, Mark Pejtersen A, Fox EA (eds) Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 1992). ACM Press, New York, pp 168–176
- Hauptmann AG, Witbrock MJ (1997) Informedia: news-on-demand multimedia information acquisition and retrieval. In: Maybury MT (ed) Intelligent multimedia information retrieval. MIT Press, Cambridge, pp 215–239
- James DA (1995) The application of classical information retrieval techniques to spoken documents. PhD thesis, Cambridge University
- Jones GJF (2000) Applying machine translation resources for cross-language information access from spoken documents. In: Proceedings of MT 2000: machine translation and multilingual applications in the new millennium. British Computer Society, pp 4-1–4-9
- Jones GJF (2001) New challenges for cross-language information retrieval: multimedia data and the user experience. In: Peters C (ed) Cross-language information retrieval and evaluation: workshop of cross-language evaluation forum (CLEF 2000). Lecture notes in computer science (LNCS), vol 2069. Springer, Heidelberg, pp 71–81
- Jones GJF (2013) An introduction to crowdsourcing for language and multimedia technology research. In: Agosti M, Ferro N, Forner P, Müller H, Santucci G (eds) Information retrieval meets information visualization – PROMISE Winter School 2012, Revised Tutorial Lectures. Lecture notes in computer science (LNCS), vol 7757. Springer, Heidelberg, pp 132–154
- Jones GJF, Federico M (2003) CLEF 2002 cross-language spoken document retrieval pilot track report. In: Peters C, Braschler M, Gonzalo J, Kluck M (eds) Advances in cross-language information retrieval: third workshop of the cross-language evaluation forum (CLEF 2002) Revised Papers. Lecture notes in computer science (LNCS), vol 2785. Springer, Heidelberg, pp 446–457
- Jones GJF, Foote JT, Spärck Jones K, Young SJ (1996) Retrieving spoken documents by combining multiple index sources. In: Frei HP, Harman D, Schaübie P, Wilkinson R (eds) Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 1996). ACM Press, New York, pp 30–38

- Kekäläinen J, Järvelin K (2002) Using graded relevance assessments in IR evaluation. *J Am Soc Inf Sci Technol* 53(13):1120–1129
- Khwhileh A, Jones GJF (2016) Investigating segment-based query expansion for user-generated spoken content retrieval. In: 14th international workshop on content-based multimedia indexing, IEEE, CBMI 2016, pp 1–6
- Khwhileh A, Afli H, Jones GJF, Way A (2017) Identifying effective translations for cross-lingual arabic-to-english user-generated speech search. In: Proceedings of the third arabic natural language processing workshop. Association for Computational Linguistics, pp 100–109
- Larson M, Jones GJF (2011) Spoken content retrieval: a survey of techniques and technologies. *Found Trends Inf Retr* 5(4–5):235–422
- Larson M, Newman E, Jones GJF (2009) Overview of VideoCLEF 2008: automatic generation of topic-based feeds for dual language audio-visual content. In: Peters C, Deselaers T, Ferro N, Gonzalo J, Jones GJF, Kurimo M, Mandl T, Peñas A (eds) Evaluating systems for multilingual and multimodal information access: ninth workshop of the cross-language evaluation forum (CLEF 2008). Revised selected papers. Lecture notes in computer science (LNCS), vol 5706. Springer, Heidelberg, pp 906–917
- Larson M, Newman E, Jones GJF (2010) Overview of VideoCLEF 2009: new perspectives on speech-based multimedia content enrichment. In: Peters C, Tsikrika T, Müller H, Kalpathy-Cramer J, Jones GJF, Gonzalo J, Caputo B (eds) Multilingual information access evaluation Vol. II multimedia experiments – tenth workshop of the cross-language evaluation forum (CLEF 2009). Revised selected papers. Lecture notes in computer science (LNCS). Springer, Heidelberg, pp 354–368
- Larson M, Eskevich M, Ordelman R, Kofler C, Schmiedeke S, Jones GJF (2011) Overview of mediaeval 2011 rich speech retrieval task and genre tagging task. In: Larson M, Rae A, Demarty CH, Kofler C, Metz F, Troncy R, Mezaris V, Jones GJF (eds) Working notes proceedings of the MediaEval 2011 multimedia benchmark workshop. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, <http://ceur-ws.org/Vol-807/>
- Marchand-Maillet S (2000) Content-based video retrieval: an overview. Technical report, Computer Vision Group, Computing Science Center, University of Geneva
- Marge M, Banerjee S, Rudnicky AI (2010) Using the Amazon Mechanical Turk for transcription of spoken language. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP 2010). IEEE, Piscataway, pp 5270–5273
- Oard DW, Wang J, Jones GJF, White RW, Pecina P, Soergel D, Huang X, Shafran I (2007) Overview of the CLEF-2006 cross-language speech retrieval track. In: Peters C, Clough P, Gey FC, Karlgren J, Magnini B, Oard DW, de Rijke M, Stempfhuber M (eds) Evaluation of multilingual and multi-modal information retrieval: seventh workshop of the cross-language evaluation forum (CLEF 2006). Revised selected papers. Lecture notes in computer science (LNCS), vol 4730. Springer, Heidelberg, pp 744–758
- Ordelman RJF, Eskevich M, Aly R, Huet B, Jones GJF (2015) Defining and evaluating video hyperlinking for navigating multimedia archives. In: Proceedings of the 24th international conference on world wide web. ACM, New York, WWW '15 Companion, pp 727–732
- Over P, Fiscus J, Joy D, Michel M, Awad G, Smeaton A, Kraaij W, Quénot G, Ordelman R, Aly R (2015) Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: The fifteenth international workshop on video retrieval evaluation (TRECVID 2015). National Institute of Standards and Technology (NIST), Special Publication 500-321, Washington
- Pecina P, Hoffmannová P, Jones GJF, Zhang Y, Oard DW (2008) Overview of the CLEF-2007 cross-language speech retrieval track. In: Peters C, Jijkoun V, Mandl T, Müller H, Oard DW, Peñas A, Petras V, Santos D (eds) Advances in multilingual and multimodal information retrieval: eighth workshop of the cross-language evaluation forum (CLEF 2007). Revised selected papers. Lecture notes in computer science (LNCS), vol 5152. Springer, Heidelberg, pp 674–686

- Racca DN, Jones GJ (2016) On the effectiveness of contextualisation techniques in spoken query spoken content retrieval. In: Perego R, Sebastiani F, Aslam J, Ruthven I, Zobel J (eds) Proceedings of the 39th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2016). ACM Press, New York, pp 933–936
- Rashchian C, Young P, Hodosh M, Hockenmaier J (2010) Collecting image annotations using Amazon’s Mechanical Turk. In: Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon’s Mechanical Turk, Association for Computational Linguistics, pp 139–147
- Sanderson M, Shou XM (2007) Search of spoken documents retrieves well recognized transcripts. In: Amati G, Carpineto C, Romano G (eds) Advances in information retrieval. Proceedings of the 29th European conference on IR research (ECIR 2007). Lecture notes in computer science (LNCS), vol 4425. Springer, Heidelberg, pp 505–516
- Schmiedeke S, Xu P, Ferrané I, Eskevich M, Kofler C, Larson M, Estève Y, Lamel L, Jones GJF, Sikora T (2013) Blip10000: a social video dataset containing SPUG content for tagging and retrieval. In: Proceedings of ACM multimedia systems. ACM, New York, MMSys’13
- Schoeffmann K, Hopfgartner F, Marques O, Böszörményi L, Jose JM (2010) Video browsing interfaces and applications: a review. *SPIE Rev I*(1):1–35
- Sheridan P, Wechsler M, Schäuble P (1997) Cross-language speech retrieval: establishing a baseline performance. In: Belkin NJ, Narasimhalu AD, Willett P, Hersh W, Can F, Voorhees EM (eds) Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 1997). ACM Press, New York, pp 99–108
- Smeaton AF, Over P, Kraaij W (2006) Evaluation campaigns and trecvid. In: Proceedings of the 8th ACM international workshop on multimedia information retrieval. ACM, New York, MIR ‘06, pp 321–330
- Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. *IEEE Trans Pattern Anal Mach Intell* 22(12):1349–1380
- White RW, Oard DW, Jones GJF, Soergel D, Huang X (2006) Overview of the CLEF-2005 cross-language speech retrieval track. In: Peters C, Gey FC, Gonzalo J, Jones GJF, Kluck M, Magnini B, Müller H, de Rijke M (eds) Accessing multilingual information repositories: sixth workshop of the cross-language evaluation forum (CLEF 2005). Revised selected papers. Lecture notes in computer science (LNCS), vol 4022. Springer, Heidelberg, pp 744–759

**Part IV**  
**Retrieval in New Domains**

# The Scholarly Impact and Strategic Intent of CLEF eHealth Labs from 2012 to 2017



Hanna Suominen, Liadh Kelly, and Lorraine Goeuriot

**Abstract** Since 2012, the CLEF eHealth initiative has aimed to gather researchers working on health text analytics and to provide them with annual shared tasks. This chapter reports on measuring its scholarly impact in 2012–2017 and describing its future objectives. The large number of submissions and citations demonstrate the substantial community interest in the tasks and their resources. Consequently, the initiative continues to run in 2018 and 2019 with its goal to support patients, their

---

**Authors' Contributions:** In alphabetical order by forename, HS, LK, and LG co-chaired the CLEF eHealth lab in 2012–2018 and led some of its tasks. To review the lab outcomes, HS first conceptualized the study and then designed and supervised its literature survey and citation content analysis which the co-authors conducted in close collaboration. HS, LK, and LG drafted the manuscript together, with dedicated duties for each coauthor. After this all authors critically commented and revised the manuscript. All authors have read and approved the final version of the chapter.

---

H. Suominen (✉)

Research School of Computer Science, The Australian National University (ANU), Canberra, ACT, Australia

Data61, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Canberra, ACT, Australia

Faculty of Science and Technology, University of Canberra, Canberra, ACT, Australia

Department of Future Technologies, University of Turku, Turku, Finland

e-mail: [hanna.suominen@anu.edu.au](mailto:hanna.suominen@anu.edu.au)

L. Kelly

Computer Science Department, Maynooth University, Maynooth, Co. Kildare, Ireland

e-mail: [liadh.kelly@mu.ie](mailto:liadh.kelly@mu.ie)

L. Goeuriot

Grenoble Informatics Laboratory, University Grenoble Alpes, Grenoble INP, LIG, Grenoble, France

e-mail: [lorraine.goeuriot@univ-grenoble-alpes.fr](mailto:lorraine.goeuriot@univ-grenoble-alpes.fr)

© Springer Nature Switzerland AG 2019

N. Ferro, C. Peters (eds.), *Information Retrieval Evaluation in a Changing World*, The Information Retrieval Series 41, [https://doi.org/10.1007/978-3-030-22948-1\\_14](https://doi.org/10.1007/978-3-030-22948-1_14)

333

family, clinical staff, health scientists, and healthcare policy makers in accessing and authoring health information in a multilingual setting.

## 1 Introduction

The requirement to ensure that patients can understand their own privacy-sensitive, official *health information* in an *Electronic Health Record (EHR)* are stipulated by policies and laws. For example, the *World Health Organization (WHO)*'s *Declaration on the Promotion of Patients' Rights in Europe* from 1994 states that all patients in healthcare services have the right to be fully informed about their own health status, conditions, prognosis, diagnoses, discharge guidelines, and proposed and alternative treatment/non-treatment with risks, benefits, and progress. It also obligates healthcare workers to give each patient a written summary of this information and communicate in a way appropriate to this patient's capacity for understanding, including minimized use of unfamiliar jargon.

Improving the readability of EHRs can contribute to patients' partial control and mastery over their own health and care, leading to their increased independence from healthcare services, better health/care decisions, and decreased costs of care (McAllister et al. 2012). This could mean replacing jargon words with patient-friendly synonyms, expanding shorthand, and an option to see the original text. The *Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT)*, *Unified Medical Language System (UMLS)*, and other terminology standards can help in defining synonym replacements and shorthand expansions, but *Natural Language Processing (NLP)* is needed to identify text snippets to be replaced with or extended by synonymous snippets. The enabling *Information Extraction (IE)* and NLP methods can also release healthcare workers' time from EHR-writing to, for example, longer discussions with the patient about the diagnosis, risks and benefits of the treatment options, and discharge guidelines.

Patient-friendly language in EHRs can help patients make informed decisions, but this also depends on their access to consumer leaflets and *other supportive information* about their health concerns in their personally-controlled EHR on the Internet. The large range of web content is widely accessible (Ilic 2010) and about 85% of people turn to its search engines for health information (Colineau and Paris 2010). EHRs can be used to naturally bridge patients' actions of reading their own EHR to searching supportive information; this *Information Retrieval (IR)* could mean enriching the EHR with hyperlinks to term definitions, care guidelines, and other information on patient-friendly and reliable sites on the Internet. Web-based EHRs that are targeted to both patients and healthcare workers for reading, writing, and sharing information are becoming increasingly common.<sup>1</sup>

---

<sup>1</sup>They have been open, for example, in Estonia (<http://www.e-tervis.ee>) and Australia (<https://myhealthrecord.gov.au>) since 2008 and 2012, respectively.



Information access conferences have organized evaluation labs on related health NLP, IE, and IR tasks for almost 20 years, as illustrated below:

- The *Text REtrieval Conference (TREC)* has considered user profiling to filter in only the topically relevant biomedical abstract in its *TREC Filtering Task* in 2000 (Robertson and Hull 2000). Its *TREC Genomics Tasks* have ranged in 2003–2007 from ad-hoc IR to text classification, passage IR, and entity-based question answering on data from biomedical papers and de-identified EHRs (Roberts et al. 2009). *TREC Medical Records Task* in 2011 targeted building a search engine where the patient cohort’s eligibility criteria for a given study can be specified through the query and then after IR on de-identified EHRs, the matching population is returned for recruiting participants (Voorhees and Tong 2011).
- Prior to *Conference and Labs of the Evaluation Forum (CLEF)* introducing its *Electronic Health (eHealth)* initiative in 2012, its *ImageCLEF* initiative organized annual *ImageCLEFmed* tasks since 2005 on image annotation, image search, and automated form filling related to image analysis tasks (Kalpathy-Cramer et al. 2011).
- In parallel in 2005–2012, the *Informatics for Integrating Biology and the Bedside* initiative has also been addressing eHealth NLP through its following shared tasks (Uzuner et al. 2011; Sun et al. 2013): text de-identification and identification of smoking status in 2006; recognition of obesity and comorbidities in 2008; medication IE in 2009; concept, assertion, and relation recognition in 2010; co-reference analysis in 2011; and temporal relations challenge in 2012.
- Also the *Medical NLP Challenges* have targeted automated diagnosis coding of radiology reports in 2007 and classifying the emotions found in suicide notes in 2011 (Pestian et al. 2011).

See Demner-Fushman and Elhadad (2016), Huang and Lu (2016), and Filannino and Uzuner (2018) for recent reviews of evaluation labs and other developments in NLP, IE, and IR of healthcare worker and patient-authored EHRs.

This chapter presents a review of CLEF eHealth outcomes in 2012–2017 and its strategic intent. The scholarly impact of the initiative is measured through the outcomes of problem specifications, resource releases, participation numbers, and citation counts. The paper extends Suominen et al. (2018a) by providing further details about the annual citation analyses and widening the scope to the future.

## 2 Materials and Methods

*Publication data* from the CLEF proceedings relevant to CLEF eHealth 2012–2017 and papers that use the CLEF eHealth datasets, based on the reference catalog of the CLEF eHealth website, were reviewed.<sup>2</sup> This *literature review* was supplemented

---

<sup>2</sup><https://sites.google.com/site/clefehealth>.

by conducting a *bibliometric study* (Tsikrika et al. 2013; Angelini et al. 2014) of the reviewed publications and their citations received by 10 Nov 2017. *Citation data* for the publication data was collected from *Google Scholar*—one of the most comprehensive citation data sources (Tsikrika et al. 2013; Angelini et al. 2014). In accordance with Tsikrika et al. (2013), we reviewed and refined the citation counts by hand for duplicated citation entries and incorrect entry merging.

*Citation content analysis* (Zhang et al. 2013a) was used for the publication and citation *data analysis*. This allowed testing of hypotheses about both the quantity and quality of the scholarly impact of CLEF eHealth in 2012–2017. See Suominen et al. (2018b) for further methodological details.

### 3 Results

Placing layperson patients to the center of these shared tasks—opposed to clinical experts—as the targeted users is the main distinguishing feature of CLEF eHealth when comparing with earlier evaluation initiatives. In 2012, CLEF eHealth ran as a *scientific workshop* with an aim of establishing an evaluation campaign (Suominen 2012b, 2014), and from 2013 to 2017 this annual workshop has been supplemented with three or more *shared tasks* each year (Fig. 1). In 2013–2017, these tasks were patient-centric, with clinicians also considered from 2015, but in 2017 a pilot task on *Technology Assisted Reviews (TAR)* to support health scientists and policymakers' information access was also introduced (Suominen et al. 2013; Kelly et al. 2014, 2016; Goeuriot et al. 2015, 2017).

#### 3.1 Problem Specifications

The first two evaluation labs, held in 2013 and 2014, focused on NLP, IR, and *Information Visualization (IV)* to support patients' in understanding their EHRs in English (see Suominen et al. (2013), Pradhan et al. (2013), Mowery et al. (2013), Goeuriot et al. (2013a) and Kelly et al. (2014), Suominen et al. (2014), Mowery et al. (2014), Goeuriot et al. (2014b) for the annual lab and Task 1–3 overviews). The *2013 Tasks 1a and 1b* considered *disorder naming* by identification of disorder names and *normalization of the identified names* by translating them to patient-friendly synonyms, respectively. The *2013 Task 2* on *shorthand expansion* aimed at mapping clinical abbreviations and acronyms to patient-friendly synonyms. Instead of actually writing the disorder names and shorthand expansions, SNOMED CT and UMLS codes were applied in Task 1b and Task 2, respectively. This challenge continued in *2014 Task 2* on *template filling*, with the aim of developing attribute classifiers that predict the *Concept Unique Identifier (CUI)* values of UMLS with mention boundaries. The Disease/Disorder Templates consisted of Negation, Uncertainty, and Severity Indicators, together with seven other attributes.

|                               | Task  | Timeline                  |   |  |                                     |                         |                                 |
|-------------------------------|---|---------------------------|---|--|-------------------------------------|-------------------------|---------------------------------|
|                               |   | 2012                      | 2013  | 2014   | 2015                                | 2016                    | 2017                            |
| <b>Information Extraction</b> | Named entity recognition and/or normalisation                 |                           | Electronic health records (EHRs) in English |  | Biomedical articles in French       |                         |                                 |
|                               | Extraction  |                           |   | EHRs in English  |                                     |                         | Multi-lingual death reports     |
|                               | Classification  |                           |   |  |                                     | Death reports in French |                                 |
|                               | Replication   |                           |   |  |                                     | Code                    | Code                            |
| <b>Information Management</b> | Visualisation   |                           |   | EHRs and other electronic health (eHealth) data in English |                                     |                         |                                 |
|                               | Report generation and management                              |                           |   |  | Nursing handover reports in English |                         |                                 |
| <b>Information Retrieval</b>  | Patient-centered information retrieval                        | Multilingual eHealth data |   |  |                                     |                         |                                 |
|                               | Cross-lingual information retrieval                           |                           | Multilingual eHealth data                   |  |                                     |                         |                                 |
|                               | Technology assisted reviews in empirical medicine             |                           |   |  |                                     |                         | Bio-medical articles in English |
| <b>Workshop</b>               |   | Yes                       | Yes   | Yes  | Yes                                 | Yes                     | Yes                             |
| <b>Participation</b>          | No. of expressions of interests to participate                |                           | 175   | 220  | 90                                  | 116                     | 117                             |
|                               | No. of participating teams                                    |                           | 34  | 24   | 20                                  | 20                      | 34                              |
|                               | No. of papers in the CLEF proceedings                         | 16                        | 34  | 29   | 24                                  | 24                      | 35                              |
|                               | No. of authors  | 50                        | 162   | 107  | 91                                  | 113                     | 128                             |
|                               | No. of authors' affiliations (academia, industry, government) | 35 (32, 2, 1)             | 85 (74, 4, 7)                               | 69 (65, 0, 4)  | 50 (47, 2, 1)                       | 73 (64, 6, 3)           | 82 (75, 5, 2)                   |
|                               | No. of affiliated countries                                   | 8                         | 10  | 22   | 19                                  | 16                      | 22                              |

**Fig. 1** Timeline of CLEF eHealth

The 2013 and 2014 Tasks 3, and 2014 Task 1 supplemented the processing of EHRs with information from the Internet, based on patient’s information needs associated with the EHRs. The 2013 and 2014 Task 3 on *information search* considered English but in 2014 the problem was extended to serving an individual expressing their information need in a non-English language, for search on web-pages written

in English because a large proportion of eHealth content on the Internet is written in English. The 2014 Task 1 on *interactive IV* had the overall goal of designing an effective, usable, and trustworthy web-environment for an English-speaking patient in their home in the USA to navigate, explore, and interpret health information as needed to promote understanding and informed decision-making.

In 2015 and 2016 the labs scope was expanded to multilingual text processing, medical web search, and speech-to-text conversion to ease both patients and clinicians' understanding of various types of medical content (see Goeuriot et al. (2015), Suominen et al. (2015a), Névéol et al. (2015), Palotti et al. (2015a) and Kelly et al. (2016), Suominen et al. (2016), Névéol et al. (2016), Zuccon et al. (2016a) for the annual lab and Task 1–3 overviews). The *2015 and 2016 Task 1* considered *nursing handover report support* in English. In clinical handover between nurses, verbal handover and note taking can lead to loss of information and electronic documentation is laborious, taking time away from patient education. The challenges addressed taking clinical notes automatically by using *Speech Recognition* to convert spoken nursing handover into digital text and *IE* to fill out a handover form, respectively. The *2015 and 2016 Task 2* considered *clinical named entity recognition* on French texts, which was previously an unexplored language. The challenges aimed to automatically identify clinically relevant entities from French biomedical articles. Also extracting causes of death from French death reports was considered. The *2015 and 2016 Task 3* considered patients' general information needs related to their medical complaints in a *cross-lingual medical search* on the web challenge. For example, their need to understand a condition or the cause of a medical symptom. The difficulty that this challenge focuses on is trying to extract relevant and reliable web pages that meet these needs expressed in English or several other languages.

The *2017 Task 1* continued the exploration of the problem of *multilingual text processing*, considering the *IE* of causes of death from both French and English death reports to ease clinicians' understanding (see Goeuriot et al. (2017), Névéol et al. (2017), Kanoulas et al. (2017), Palotti et al. (2017) for the annual lab and Task 1–3 overviews). The *2017 Task 3* also continued its exploration of developing *medical web search* techniques to address the challenge posed by patients in locating relevant and reliable medical content. In addition the *2017 Task 2* considered a new challenge, that of *TAR generation* in empirical medicine to support health care and policy making. Medical researchers and policy-makers while writing systematic review articles must ensure that they consider all documents relevant to their review. As the amount of medical literature continues to expand, automation in this process is necessary.

## 3.2 Data Releases

In 2013 Task 1, the de-identified, annotated EHRs were part of the *Shared Annotated Resources (ShARe)* corpus of the *Multiparameter Intelligent Monitoring in Intensive*

Care (MIMIC) II database.<sup>3</sup> These 300 EHRs were authored in US intensive care. Each EHR was annotated by two people. A disorder name was defined as any text snippet which fulfills the following three conditions: (1) The snippet can be mapped to a concept in SNOMED CT. (2) This concept belongs to the semantic group of Disorder. (3) The concept belongs to one of the following semantic types in UMLS: Acquired abnormality, Anatomical abnormality, Cell or molecular dysfunction, Congenital Abnormality, Disease or Syndrome, Experimental Model of Disease, Injury or Poisoning, Mental or Behavioral Dysfunction, Neoplastic Process, Pathologic Function, or Signs and Symptoms. The same EHRs and annotations were used for 2013 Tasks 1b and supplemented by a new annotation for Task 2. Thirteen people were trained for the task and provided the visually annotated Task 1 EHRs. They were instructed to mark and, when possible, codify each clinical shorthand in the EHRs with one UMLS CUI or assign the code CUI-less.

An option to use the Task 1 and 2 EHRs and annotations for 2013 Task 3 was given but to enable IR, 55 new search tasks were formed (Goeuriot et al. 2013b). Two people formed the tasks from the Task 1 materials. For each search task, they generated a *patient profile*, *information need*, *query title*, and *query description*. The profile also allowed the participants to address the task without considering the EHRs. To create result document sets for these search tasks, about one million documents from commonly used health and medicine web-sites were used (Hanbury and Müller 2012). The relevance of each document was assessed by one person.

For the 2014 Task 1 on IV, six patient cases were chosen from these 2013 Tasks 1–3 data. After the task, the workspace was kept open for registration; by 26 Oct 2017, access had been granted to 60 people.<sup>4</sup>

The 2014 Task 2 on template filling used the original 300 EHRs from 2013 Task 1 and unseen 133 EHRs. The 2013 annotations were extended by focusing on the attributes-template filling for each disorder mention. Each EHR was annotated by two people. For 2014 Task 3, two people created 55 queries from the main disorders diagnosed in these EHRs. The 2013 document collection was used and associated result sets for the queries generated. The relevance of each document was assessed by one person. Participants were provided with the mapping between queries and EHRs, and were free to use the EHRs.

For 2015 Task 3 on IR, web-documents of the 2013 Task 3 were used. Queries were obtained by showing images and videos related to medical symptoms to users, who were then asked which queries they would issue to a web search engine if they were exhibiting such symptoms and thus wanted to find more information to understand these symptoms or which condition they were affected by. Twelve people generated the queries. A total of 266 unique queries were collected; of these, 67 queries in English were selected to be used in the task. The queries' translation was also provided into Arabic, Czech, German, Farsi, French, Italian, and Portuguese. Relevance and readability assessments were performed by four people.

---

<sup>3</sup><https://www.clinicalnlpannotation.org>, <http://mimic.physionet.org>.

<sup>4</sup><https://physionet.org/works/CLEFeHealth2014Task1/>.

The 2016 Task 3 on IR, used a new corpus, *ClueWeb12 B13*,<sup>5</sup> which is a large snapshot of the web (approx. 52.3 million web pages), crawled in Feb–May 2012. Unlike the dataset used in 2013–2015 IR Tasks, the corpus did not contain only health-related pages, making the dataset more in line with the material current web search engines index and retrieve. The queries extended upon the focus of the 2015 Task 3 (self-diagnosis) by considering real health information needs expressed by the general public through posts published in public health web forums. Forum posts were extracted from the ‘askDocs’ section of *Reddit*<sup>6</sup> and presented to six people, who were asked to formulate English queries based on what they read in the initial user post. This led to a set of query variants for a fixed number of topics. For the query variations element of the task, participants were told which queries relate to the same information need, to allow them to produce one set of results to be used as answer for all query variations of an information need. For the multilingual element of the task, Czech, French, German, Hungarian, Polish, and Swedish translations of the queries were provided. People assessed the outcomes for relevance, readability, and reliability. The 2017 Task 3 used the document collection and topics of 2016 Task 3, with the aim to acquire more relevance assessments and improve the collection re-usability.

The 2015 Task 1 and 2016 Task 1 on nursing handover report support used the *NICTA Synthetic Nursing Handover Data* (Suominen et al. 2015b). This set of 300 synthetic patient cases was developed for speech recognition and IE related to nursing shift-change handover. Each case was authored by a registered nurse and consisted of a patient profile; a written, free-form text paragraph to be used as a reference standard in speech recognition; its spoken and speech-recognized counterparts; and human-annotations with respect to a form with 49 headings to fill out.

For 2015 Task 2 on IE, two types of biomedical documents were used: a total of 1668 titles of scientific articles indexed in the *MEDLINE* database, and six full text drug monographs published by *European Medicines Agency (EMA)*. Annotations covered ten types of entities of clinical interest, defined by ten UMLS Semantic Groups. Three people marked each relevant entity mention in the documents, and assigned the corresponding semantic types and CUIs (Névéol et al. 2014). The 2016 Task 2 extended this 2015 Task 2 data release by including 833 *MEDLINE* titles and 4 *EMA* documents, with annotations for ten types of clinical entities with UMLS normalization. In another challenge, it used 65,843 death certificates from the *CépiDC Causes of Death Corpus* that were manually coded with *International Classification of Diseases (ICD)-10*, as per the WHO standards. The 2017 Task 1 supplemented these French death certificates by those in English from the USA. The annotators at the *French National Institute of Health and Medical Research (INSERM)* in 2006–2013 and the US Center for Disease Control in 2015 also

---

<sup>5</sup><http://lemurproject.org/clueweb12/index.php>.

<sup>6</sup><https://www.reddit.com/r/AskDocs/>.

manually built dictionaries of terms associated with the codes. Several versions of these lexical resources were supplied to participants.

The new TAR in empirical medicine task—2017 Task 2—used a subset of PubMed documents for its challenge to make Abstract and Title Screening more effective when judging whether to include/exclude a reference or consider it for further examination at a full content level. The PubMed document IDs were collected by the task coordinators by re-running the MEDLINE Boolean query used in the original systematic reviews conducted by Cochrane to search 50 topics.

### 3.3 *Software Releases*

With an aim to lower the entry barrier and encourage novelty in problem solutions, CLEF eHealth began providing participants with software and code in 2013. These resource releases targeted method evaluation, EHR text annotation, and document relevance assessment, as illustrated below.

First, in 2013 Tasks 1 and 2, we released both a command-line tool and a graphical user interface that the participants could use to compute the values for the official and supplementary evaluation measures and visualize annotations against their method outputs. This *extensible Human Oracle Suite of Tools*<sup>7</sup> (South et al. 2014) also supported them in annotating more data.

Second, in 2013 Task 3, we released the *Relevation! tool*<sup>8</sup> (Koopman and Zucon 2014). We also provided a pointer to an established tool for computing values for the official and supplementary evaluation measures.

Third, 2016 Task 1 released the organizers' entire software stack as a state-of-the-art solution to the handover problem (Suominen et al. 2015b).<sup>9</sup> Participants were welcomed to use the released code for feature generation and/or IE, as intended, the results highlighted all participating teams' methods outperforming this known state-of-the-art baseline.

### 3.4 *Papers and Their Citations*

In 2012, the CLEF initiative introduced eHealth as a workshop that focused on eHealth documents and related analytics with a goal to spin out an evaluation lab. Its program consisted of three invited talks on collaborative datasets, resources, tools, and infrastructure; an expert panel; a student mentoring session where champions of the field provided feedback on designated PhD study plans and projects; a

---

<sup>7</sup><http://blulab.chpc.utah.edu/content/ehost-extensible-human-oracle-suite-tools>.

<sup>8</sup><http://ielab.github.io/relevation>.

<sup>9</sup><https://www.kaggle.com/c/hospital-handover-forms/>.

**Table 1** Bibliometric analysis of the CLEF eHealth 2012 lab workshop on 26 Oct 2017

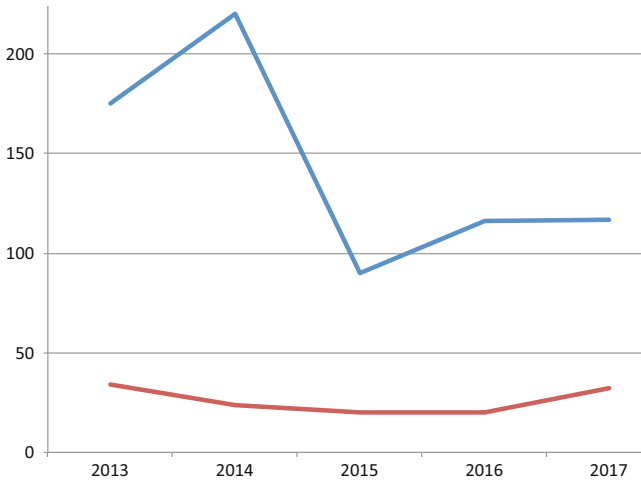
| ID                    | Paper                              | Authors | Authors' countries                             | Citations |
|-----------------------|------------------------------------|---------|--|-----------|
| Organizers' overview: |                                    |         |  |           |
| 1                     | Suominen (2012a)                   | 1       | Australia                                      | 2         |
| Participants' papers: |                                    |         |  |           |
| 2                     | Friberg Heppin and Järvelin (2012) | 2       | Sweden   | 1         |
| 3                     | Heimonen et al. (2012)             | 3       | Finland  | 0         |
| 4                     | Isenius et al. (2012)              | 3       | Sweden   | 10        |
| 5                     | Kanhov et al. (2012)               | 3       | Sweden   | 0         |
| 6                     | Kelly et al. (2012)                | 4       | Austria, Ireland                               | 1         |
| 7                     | Kreiner et al. (2012)              | 4       | Austria  | 0         |
| 8                     | Laippala et al. (2012)             | 5       | Finland  | 0         |
| 9                     | Martinez et al. (2012)             | 4       | Australia                                      | 5         |
| 10                    | Moen and Marsi (2012)              | 2       | Norway   | 0         |
| 11                    | Suominen et al. (2012a)            | 8       | Australia                                      | 1         |
| 12                    | Suominen et al. (2012b)            | 3       | Australia, Austria                             | 0         |
| Invited abstracts:    |                                    |         |  |           |
| 13                    | Chapman (2012)                     | 1       | USA  | 0         |
| 14                    | Hanlen (2012)                      | 1       | Australia                                      | 0         |
| 15                    | Jones et al. (2012)                | 5       | Austria, Finland, Ireland, Sweden, Switzerland | 0         |
| 16                    | Suominen (2012b)                   | 1       | Australia                                      | 0         |

professional networking session; a working session for developing a road map for CLEF eHealth 2013; and oral talks for eleven papers (Table 1).

All CLEF eHealth 2012 talks focused on meeting the needs of healthcare professionals and patients in ease of information recording, access, and understanding via user-centered abbreviation processing, content generation, search engines, and vocabularies, among other tools to support patient–professional interaction across languages, sub-languages, and jargons. This community interest in the topic of user-friendly multilingual communication was verified in the roadmap session and formed the focus of the successful CLEF eHealth 2013–2017 labs (Figs. 1 and 2).

From 2012 to 2017, the 184 CLEF eHealth papers with 1299 citations generated in total the scholarly citation impact of almost  $741 \times 1299 = 963,000$  citations for the 741 co-authors and reached authors from 33 countries across the world (Tables 1, 2, 3, 4, 5, 6, and 7). In accordance with the CLEF eHealth mission to foster teamwork, the number of co-authors per paper was 4 on average, with the maximum, median, minimum, and standard deviation of 15, 3, 1, and 3, respectively. In 47 out of the 184 papers (26%), this co-authoring collaboration was international.





**Fig. 2** Participation (red) and expression of interest (blue) in the CLEF eHealth evaluation labs

143 out of the 184 papers (78%) had been cited at least once. The number of citations per paper varied from 0 to 147, with the mean and standard deviation of 7 and 15, respectively. The *h-index* and *i10-index* were 18 and 35, respectively. In 2012 CLEF eHealth resulted in 16 papers and in 2013–2017, this number increased to 31–35.

## 4 Discussion

CLEF eHealth has been contributing to evaluation initiatives in medical NLP and IR since 2012. Evaluation resources have been developed and shared with the community to support the understanding of and access to health content by laypeople, clinicians, and policy-makers. In so doing the lab has provided an evaluation setting for the progression of multilingual eHealth *Information and Communications Technology (ICT)* research and development. The annual eHealth workshop held at the main CLEF conference provides for the dissemination and discussion of the outcomes of each year’s tasks. Each year the organizers produce overview papers describing the tasks offered and participants results. These have proven influential, as indicated by their citation indexes.

Although the CLEF eHealth installations have attracted substantial community interest, as reflected by the 741 co-authors of the 184 papers from 33 countries, substantially more participation from Central America, Africa, South America, and the Middle East should be achievable. However, this problem of insufficient participation has been acknowledged in a review of biomedical evaluation initiatives by Huang and Lu (2016) as one of their main conclusions.

**Table 2** Bibliometric analysis of the CLEF eHealth 2013 evaluation lab on 31 Oct 2017

| ID  | Paper                               | Authors | Authors' countries                                | Citations |
|---|-------------------------------------|---------|---|-----------|
| Organizers' overviews:                      |                                     |         |   |           |
| 1   | Suominen et al. (2013)              | 15      | Australia, Finland, Ireland, Sweden, USA          | 147       |
| 2   | Pradhan et al. (2013)               | 9       | Australia, USA                                    | 29        |
| 3   | Mowery et al. (2013)                | 11      | Australia, Finland, USA                           | 21        |
| 4   | Goeriot et al. (2013a)              | 9       | Australia, Austria, Finland, Ireland, Switzerland | 48        |
| Participants' papers for 2013 Task 1:       |                                     |         |   |           |
| 5   | Bodnari et al. (2013)               | 5       | France, USA                                       | 17        |
| 6   | Cogley et al. (2013)                | 3       | Ireland   | 10        |
| 7   | Fan et al. (2013)                   | 3       | USA   | 4         |
| 8   | Gung (2013)                         | 1       | USA   | 8         |
| 9   | Hervas et al. (2013a)               | 4       | Spain   | 3         |
| 10  | Hervas et al. (2013b)               | 4       | Spain   | 3         |
| 11  | Leaman et al. (2013)                | 3       | USA   | 29        |
| 12  | Liu et al. (2013)                   | 4       | USA   | 8         |
| 13  | Osborne et al. (2013)               | 3       | USA   | 14        |
| 14  | Patrick et al. (2013a)              | 3       | Australia   | 5         |
| 15  | Ramanan et al. (2013)               | 3       | India   | 5         |
| 16  | Tang et al. (2013)                  | 5       | China, USA  | 18        |
| 17  | Wang and Akella (2013)              | 2       | USA   | 5         |
| 18  | Xia et al. (2013a)                  | 7       | China   | 10        |
| 19  | Zuccon et al. (2013a)               | 4       | Australia   | 6         |
| Participants' papers for 2013 Task 2:       |                                     |         |   |           |
| 20  | Jagannathan et al. (2013)           | 7       | USA   | 1         |
| 21  | Patrick et al. (2013b)              | 3       | Australia   | 5         |
| 22  | Wu et al. (2013)                    | 6       | USA   | 6         |
| 23  | Xia et al. (2013b)                  | 7       | China   | 3         |
| 24  | Zweigenbaum et al. (2013)           | 5       | France  | 0         |
| Participants' papers for 2013 Task 3:       |                                     |         |   |           |
| 25  | Bedrick and Sheikshabbafghi (2013)  | 2       | USA   | 4         |
| 26  | Caballero Barajas and Akella (2013) | 2       | USA   | 8         |
| 27  | Chappell and Geva (2013)            | 2       | Australia   | 2         |
| 28  | Choi and Choi (2013)                | 2       | Republic of Korea                                 | 5         |
| 29  | Limsopatham et al. (2013)           | 3       | UK  | 2         |
| 30  | Zhang et al. (2013b)                | 5       | USA   | 2         |
| 31  | Zhong et al. (2013)                 | 6       | China   | 5         |
| 32  | Zhu et al. (2013)                   | 5       | USA   | 20        |
| 33  | Zuccon et al. (2013b)               | 3       | Australia   | 5         |
| Participants' papers for student mentoring: |                                     |         |   |           |
| 34  | Murtola et al. (2013)               | 6       | Finland, India, Norway                            | 0         |

**Table 3** Bibliometric analysis of the CLEF eHealth 2014 evaluation lab on 9 Nov 2017

| ID                                    | Paper                               | Authors | Authors' countries                                       | Citations |
|---------------------------------------|-------------------------------------|---------|--|-----------|
| Organizers' overviews:                |                                     |         |  |           |
| 1                                     | Kelly et al. (2014)                 | 11      | Australia, Austria, Germany, Ireland, Sweden, USA        | 70        |
| 2                                     | Suominen et al. (2014)              | 10      | Australia, Germany, Ireland, USA                         | 7         |
| 3                                     | Mowery et al. (2014)                | 11      | Australia, Ireland, Sweden, USA                          | 21        |
| 4                                     | Goeuriot et al. (2014b)             | 9       | Australia, Austria, Czech Republic, Ireland, Switzerland | 65        |
| Participants' papers for 2014 Task 1: |                                     |         |  |           |
| 5                                     | Hyman and Fridy (2014)              | 2       | USA  | 1         |
| Participants' papers for 2014 Task 2: |                                     |         |  |           |
| 6                                     | Hamon et al. (2014)                 | 3       | France   | 5         |
| 7                                     | Herbst et al. (2014)                | 4       | Germany  | 3         |
| 8                                     | Huynh and Ho (2014)                 | 2       | Vietnam  | 4         |
| 9                                     | Johri et al. (2014)                 | 3       | India, Japan   | 6         |
| 10                                    | Liu and Ku (2014)                   | 2       | Taiwan   | 1         |
| 11                                    | Liu et al. (2014)                   | 4       | Canada   | 0         |
| 12                                    | Mkrtchyan and Sonntag (2014)        | 2       | Germany  | 5         |
| 13                                    | Osborne (2014)                      | 1       | USA  | 0         |
| 14                                    | Ramanan and Senthil Nathan (2014)   | 2       | India  | 3         |
| 15                                    | Sequeira et al. (2014)              | 4       | Portugal   | 2         |
| Participants' papers for 2014 Task 3: |                                     |         |  |           |
| 16                                    | Choi and Choi (2014)                | 2       | Republic of Korea  | 12        |
| 17                                    | Claveau et al. (2014)               | 4       | France   | 3         |
| 18                                    | Dramé et al. (2014)                 | 3       | France   | 11        |
| 19                                    | Malagon and L'opez (2014)           | 2       | Spain  | 0         |
| 20                                    | Nesrine et al. (2014)               | 3       | Tunisia  | 4         |
| 21                                    | Oh and Jung (2014)                  | 2       | Republic of Korea  | 9         |
| 22                                    | Ozturkmenoglu et al. (2014)         | 3       | Turkey   | 4         |
| 23                                    | Saleh and Pecina (2014)             | 2       | Czech Republic   | 4         |
| 24                                    | Shenwei et al. (2014)               | 4       | Canada   | 15        |
| 25                                    | Thakkar et al. (2014)               | 4       | India  | 8         |
| 26                                    | Thesprasith and Jaruskulchai (2014) | 2       | Thailand   | 3         |
| 27                                    | Verberne (2014)                     | 1       | Netherlands  | 3         |
| 28                                    | Wu and Huang (2014)                 | 2       | Canada   | 0         |
| 29                                    | Yang et al. (2014)                  | 3       | USA  | 4         |

**Table 4** Bibliometric analysis of the CLEF eHealth 2015 evaluation lab on 10 Nov 2017

| ID                                    | Paper                                | Authors | Authors' countries                                  | Citations |
|---------------------------------------|--------------------------------------|---------|---|-----------|
| Organizers' overviews:                |                                      |         |   |           |
| 1                                     | Goeuriot et al. (2015)               | 8       | Australia, Austria, France, Ireland                 | 37        |
| 2                                     | Suominen et al. (2015a)              | 5       | Australia, France, Ireland                          | 1         |
| 3                                     | Névéol et al. (2015)                 | 7       | France  | 17        |
| 4                                     | Palotti et al. (2015a)               | 8       | Australia, Austria, Czech Republic, France, Ireland | 38        |
| Participants' papers for 2015 Task 1: |                                      |         |   |           |
| 5                                     | Herms et al. (2015)                  | 4       | Germany   | 3         |
| 6                                     | Luu et al. (2015)                    | 4       | Australia   | 1         |
| Participants' papers for 2015 Task 2: |                                      |         |   |           |
| 7                                     | Afzal et al. (2015)                  | 5       | Netherlands   | 4         |
| 8                                     | Chernyshevich and Stankevitch (2015) | 2       | Belarus   | 4         |
| 9                                     | Cotik et al. (2015)                  | 3       | Argentina, Spain                                    | 1         |
| 10                                    | D'Hondt et al. (2015b)               | 6       | France  | 3         |
| 11                                    | Jain (2015)                          | 1       | India   | 3         |
| 12                                    | Jiang et al. (2015)                  | 3       | China   | 3         |
| 13                                    | Soualmia et al. (2015)               | 4       | France  | 3         |
| Participants' papers for 2015 Task 3: |                                      |         |   |           |
| 13                                    | As above                             |         |   |           |
| 14                                    | D'Hondt et al. (2015a)               | 3       | France  | 0         |
| 15                                    | Ghoddousi and Huang (2015)           | 2       | Canada  | 0         |
| 16                                    | Huynh et al. (2015)                  | 3       | Vietnam   | 2         |
| 17                                    | Ksentini et al. (2015)               | 4       | France, Tunisia                                     | 2         |
| 18                                    | Liu and Nie (2015)                   | 2       | Canada  | 3         |
| 19                                    | Lu (2015)                            | 1       | China   | 2         |
| 20                                    | Oh et al. (2015)                     | 3       | Republic of Korea                                   | 2         |
| 21                                    | Saleh et al. (2015)                  | 3       | Czech Republic                                      | 1         |
| 22                                    | Song et al. (2015)                   | 5       | China, USA  | 5         |
| 23                                    | Thesprasith and Jaruskulchai (2015)  | 2       | Thailand  | 1         |
| 24                                    | Thuma et al. (2015)                  | 3       | Botswana  | 2         |

**Table 5** Bibliometric analysis of the CLEF eHealth 2016 evaluation lab on 10 Nov 2017

| ID                                    | Paper                           | Authors | Authors' countries   | Citations |
|---------------------------------------|---------------------------------|---------|--|-----------|
| Organizers' overviews:                |                                 |         |  |           |
| 1                                     | Kelly et al. (2016)             | 6       | Australia, Austria, France, Ireland                              | 21        |
| 2                                     | Suominen et al. (2016)          | 4       | Australia, France, Ireland                                       | 5         |
| 3                                     | Névéol et al. (2016)            | 11      | France, Ireland  | 18        |
| 4                                     | Zuccon et al. (2016a)           | 9       | Australia, Austria, Czech Republic, France, Ireland, Switzerland | 23        |
| Participants' papers for 2016 Task 1: |                                 |         |  |           |
| 5                                     | Ebersbach et al. (2016)         | 4       | Germany  | 3         |
| 6                                     | Quiroz et al. (2016)            | 4       | Netherlands  | 2         |
| 7                                     | Song et al. (2016a)             | 6       | China  | 2         |
| Participants' papers for 2016 Task 2: |                                 |         |  |           |
| 8                                     | Cabot et al. (2016)             | 4       | France   | 8         |
| 9                                     | Dermouche et al. (2016)         | 6       | France   | 10        |
| 10                                    | Ho-Dac et al. (2016)            | 8       | France   | 2         |
| 11                                    | Mottin et al. (2016)            | 6       | Switzerland  | 3         |
| 12                                    | Saleh and Pecina (2016)         | 2       | Czech Republic   | 1         |
| 13                                    | van Mulligen et al. (2016)      | 5       | Netherlands  | 0         |
| 14                                    | Vivaldi et al. (2016)           | 3       | Argentina, Spain   | 1         |
| 15                                    | Zweigenbaum and Lavergne (2016) | 8       | France   | 7         |
| Participants' papers for 2016 Task 3: |                                 |         |  |           |
| 16                                    | Budaher et al. (2016)           | 3       | France   | 0         |
| 17                                    | Oh and Jung (2016)              | 2       | Republic of Korea  | 0         |
| 18                                    | Silva and Lopes (2016)          | 2       | Portugal   | 0         |
| 19                                    | Soldaini et al. (2016)          | 3       | USA  | 1         |
| 20                                    | Song et al. (2016b)             | 6       | China  | 1         |
| 21                                    | Thuma et al. (2016)             | 3       | Botswana   | 1         |
| 22                                    | Ullah and Aono (2016)           | 2       | Japan  | 0         |
| 23                                    | Wang et al. (2016a)             | 3       | China  | 0         |
| 24                                    | Wang et al. (2016b)             | 3       | USA  | 1         |

**Table 6** Bibliometric analysis of the CLEF eHealth 2017 evaluation lab on 10 Nov 2017

| ID                                    | Paper                                     | Authors | Authors' countries                                  | Citations |
|---------------------------------------|---|---------|---|-----------|
| Organizers' overviews:                |   |         |   |           |
| 1                                     | Goeuriot et al. (2017)                    | 9       | Australia, Austria, France, Ireland, Netherlands    | 22        |
| 2                                     | Névéal et al. (2017)                      | 9       | France, USA   | 10        |
| 3                                     | Kanoulas et al. (2017)                    | 4       | Netherlands, UK                                     | 10        |
| 4                                     | Palotti et al. (2017)                     | 8       | Australia, Austria, Czech Republic, France, Ireland | 4         |
| Participants' papers for 2017 Task 1: |   |         |   |           |
| 5                                     | Atemezing (2017)                          | 1       | France  | 0         |
| 6                                     | Cabot et al. (2017)                       | 3       | France  | 1         |
| 7                                     | Di Nunzio et al. (2017b)                  | 4       | Italy   | 0         |
| 8                                     | Ebersbach et al. (2017)                   | 3       | Germany   | 1         |
| 9                                     | Ho-Dac et al. (2017)                      | 12      | France  | 1         |
| 10                                    | Jonnagaddala and Hu (2017)                | 2       | Australia, Ireland                                  | 1         |
| 11                                    | Miftahutdinov and Tutubalina (2017)       | 2       | Russia  | 0         |
| 12                                    | Seva et al. (2017)                        | 4       | Germany   | 1         |
| 13                                    | Tchechmedjiev et al. (2017)               | 4       | France, USA   | 0         |
| 14                                    | Zweigenbaum and Lavergne (2017)           | 2       | France  | 2         |
| Participants' papers for 2017 Task 2: |   |         |   |           |
| 15                                    | Alharbi and Stevenson (2017)              | 2       | UK  | 0         |
| 16                                    | Anagnostou et al. (2017)                  | 4       | Greece  | 0         |
| 17                                    | Azzopardi et al. (2017)                   | 3       | UK  | 2         |
| 18                                    | Chen et al. (2017)                        | 7       | China   | 2         |
| 19                                    | Cormack and Grossman (2017)               | 2       | Canada  | 2         |
| 20                                    | Di Nunzio et al. (2017a)                  | 4       | Italy   | 2         |
| 21                                    | Lee (2017)                                | 1       | Singapore   | 0         |
| 22                                    | Norman et al. (2017)                      | 3       | France, Netherlands                                 | 2         |
| 23                                    | Scells et al. (2017)                      | 4       | Australia   | 2         |
| 24                                    | Singh and Thomas (2017)                   | 2       | India   | 0         |
| 25                                    | Singh et al. (2017)                       | 4       | UK, USA   | 2         |
| 26                                    | van Altena and Delgado Olabarriaga (2017) | 2       | Netherlands   | 2         |
| 27                                    | Yu and Menzies (2017)                     | 2       | USA   | 0         |
| Participants' papers for 2017 Task 3: |   |         |   |           |
| 28                                    | Diaz-Galiano et al. (2017)                | 5       | Spain   | 0         |
| 29                                    | Hollmann and Eickhoff (2017)              | 2       | Switzerland   | 0         |
| 30                                    | Jimmy et al. (2017)                       | 3       | Australia, Indonesia                                | 0         |
| 31                                    | Oh and Jung (2017)                        | 2       | Republic of Korea                                   | 0         |
| 32                                    | Palotti and Rekabsaz (2017)               | 2       | Austria   | 0         |
| 33                                    | Saleh and Pecina (2017)                   | 2       | Czech Republic                                      | 1         |
| 34                                    | Thuma et al. (2017)                       | 3       | Botswana  | 0         |
| 35                                    | Yang and Goncalves (2017)                 | 2       | Portugal  | 0         |

**Table 7** Bibliometric analysis of other papers that use CLEF eHealth data on 10 Nov 2017

| ID | Paper                          | Authors | Authors' countries                       | Citations |
|----|--------------------------------|---------|--|-----------|
| 1  | Goeriot et al. (2013b)         | 8       | Australia, Austria, Ireland, Switzerland | 11        |
| 2  | Goeriot et al. (2014a)         | 3       | Ireland                                  | 6         |
| 3  | Kholghi et al. (2014)          | 4       | Australia                                | 5         |
| 4  | Pradhan et al. (2015)          | 9       | Australia, USA                           | 53        |
| 5  | Suominen (2014)                | 1       | Australia, Finland                       | 4         |
| 6  | Zuccon and Koopman (2014)      | 2       | Australia                                | 22        |
| 7  | De Vine et al. (2015)          | 4       | Australia                                | 5         |
| 8  | Kholghi et al. (2015)          | 4       | Australia                                | 5         |
| 9  | Palotti et al. (2015b)         | 3       | Australia, Austria                       | 6         |
| 10 | Suominen et al. (2015b)        | 4       | Australia, Finland                       | 15        |
| 11 | Zhou and Suominen (2015)       | 2       | Australia, Finland                       | 0         |
| 12 | Zhou et al. (2015)             | 3       | Australia, Finland                       | 2         |
| 13 | Zuccon et al. (2015)           | 3       | Australia, Austria                       | 21        |
| 14 | Beloborodov and Goeriot (2016) | 2       | France, Russia                           | 6         |
| 15 | Goeriot et al. (2016)          | 4       | Australia, Austria, France, Ireland      | 2         |
| 16 | Kholghi et al. (2016)          | 4       | Australia                                | 7         |
| 17 | Mowery et al. (2016)           | 13      | Australia, Finland, Sweden, USA          | 1         |
| 18 | Palotti et al. (2016a)         | 4       | Australia, Austria, France               | 6         |
| 19 | Palotti et al. (2016b)         | 5       | Australia, Austria, France               | 21        |
| 20 | Rekabsaz et al. (2016)         | 4       | Australia, Austria                       | 7         |
| 21 | Zuccon (2016)                  | 1       | Australia                                | 17        |
| 22 | Zuccon et al. (2016b)          | 3       | Australia, Austria                       | 8         |

By virtue of the lab series over the first 7 years of its life, from 2012 to 2018 inclusive, providing access to shared data, resources, processing methods, and evaluation settings for medical system research, development and evaluation; offering reproducibility, scalability, and user-centricity; and finally bringing the research community together through the lab series to collaborate and discuss challenges associated with technique development in medical NLP and IR, we conjecture that CLEF eHealth has impacted progress in these spaces. While it is difficult to accurately quantify such impact, the 1299 citations, with impact of circa 963,000 generated by the lab in its first 6 year's of existence are suggestive. Progress in the areas addressed by the lab has the potential to generate high impact not only on the research field, but more generally on society, given the importance of health information access to support healthcare as well as to empower people to manage their health. Consequently, CLEF eHealth runs in 2018 and 2019 extending its previous challenges (Suominen et al. 2018b).

Going forward, the strategic intent of the CLEF eHealth initiative is to develop shared tasks that influence the patient care continuum by impacting (1) patient understanding of their health and healthcare, and (2) the entire healthcare ecosystem

which exists to support patient care. To achieve this, we continue to provide the community with an increasingly sophisticated dataset of clinical narrative, enriched with links to evidence-based care guidelines, systematic reviews, and other further information, to advance the state-of-the-art in multilingual NLP and IR in healthcare. Our scope fosters student mentoring, diverse collaboration, and reproducible research by welcoming and supporting new participants; facilitating multi-professional and interdisciplinary collaboration; and encouraging participants to reflect on methods and practical steps to take to facilitate the replication of their experiments; fostering the release of open-source datasets and tools to reach a wider community. This scope is supported by an increasing interest of the community in health-related IR and NLP, and its increased consideration for shared tasks.

**Acknowledgements** We gratefully acknowledge the people involved in CLEF eHealth 2012–2017 as participants or organizers. We are also thankful to support by the CLEF Initiative; Data61; *European Science Foundation (ESF) Project Evaluating Information Access Systems (ELIAS)* network program Horizon 2020 program (H2020-ICT-2014-1) under grant agreement 644753 (KConnect); *French National Research Agency (ANR)*, under grant CBeRneT ANR-13-JS02-0009-01; *Knowledge Helper for Medical and Other Information users (Khresmoi)* project, funded by the European Union *Seventh Framework Programme (FP7)*/2007–2013 under grant agreement No. 257528; Microsoft Azure for Research Award CRM:0518649; MIMIC II Database; *National Information and Communications Technologies Australia (NICTA)*, funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program; Protégé resource, supported by grant GM10331601 from the National Institute of General Medical Sciences of the *United States (of America) (US)* National Institutes of Health; PhysioNetWorks Workspaces; ShARe project funded by the US National Institutes of Health (R01GM090187); Swedish Research Council (350-2012-6658); Swedish Vårdal Foundation; US Department of *Veterans Affairs (VA)* Consortium for *Healthcare Informatics Research (CHIR)*; and US Office of the National Coordinator of Healthcare Technology, *Strategic Health Information Technology Advanced Research Projects (SHARP)* 90TR0002.

## References

- Afzal Z, Akhondi S, van Haagen H, van Mulligen E, Kors J (2015) Biomedical concept recognition in French text using automatic translation of English terms. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Alharbi A, Stevenson M (2017) Ranking abstracts to identify relevant evidence for systematic reviews: The University of Sheffield's approach to CLEF eHealth 2017 task 2. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Anagnostou A, Lagopoulos A, Tsoumakas G, Vlahavas I (2017) Combining inter-review learning-to-rank and intra-review incremental training for title and abstract screening in systematic reviews. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Angelini M, Ferro N, Larsen B, Müller H, Santucci G, Silvello G, Tsikrika T (2014) Measuring and analyzing the scholarly impact of experimental evaluation initiatives. *Procedia Comput Sci* 38(Supplement C):133–137



- Atemezing G (2017) NoNLP: annotating medical domain by using semantic technologies. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Azzopardi L, Kalphov V, Georgiadis G (2017) SiS at CLEF 2017 eHealth TAR task. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Bedrick S, Sheikshabbafghi G (2013) Lucene, MetaMap, and language modeling: OHSU at CLEF eHealth 2013. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Beloborodov A, Goeuriot L (2016) Improving health consumer search with contextual information. In: Proceedings of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR). Association for Computing Machinery, New York
- Bodnari A, Deleger L, Lavergne T, Neveol A, Zweigenbaum P (2013) A supervised named-entity extraction system for medical text. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Budaheer J, Almasri M, Goeuriot L (2016) Comparison of several word embedding sources for medical information retrieval. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Caballero Barajas K, Akella R (2013) Incorporating statistical topic models in the retrieval of healthcare documents. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Cabot C, Soualmia L, Dahamna B, Darmoni S (2016) SIBM at CLEF eHealth evaluation lab 2016: extracting concepts in French medical texts with ECMT and CIMIND. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Cabot C, Soualmia L, Darmoni S (2017) SIBM at CLEF eHealth evaluation lab 2017: multilingual information extraction with CIM-IND. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Chapman W (2012) Developing resources to assist in development and application of NLP to clinical texts. In: CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Chappell T, Geva S (2013) Working notes for TopSig at ShARe/CLEF eHealth 2013. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Chen J, Chen S, Song Y, Liu H, Wang Y, Hu Q, He L (2017) ECNU at 2017 eHealth task 2: technologically assisted reviews in empirical medicine. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Chernyshevich M, Stankevitch V (2015) IHS-RD-BELARUS: clinical named entities identification in French medical texts. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Choi S, Choi J (2013) SNUMedinfo at CLEFeHealth2013 task 3. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Choi S, Choi J (2014) Exploring effective information retrieval technique for the medical web documents: SNUMedinfo at CLEFeHealth2014 task 3. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Claveau V, Hamon T, Grabar N, Le Maguer S (2014) RePaLi participation to CLEF eHealth IR challenge 2014: leveraging term variation. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Cogley J, Stokes N, Carthy J (2013) Medical disorder recognition with structural support vector machines. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Colineau N, Paris C (2010) Talking about your health to strangers: understanding the use of online social networks by patients. *New Rev Hypermedia Multimedia* 16(1–2):141–160

- Cormack G, Grossman M (2017) Technology-assisted review in empirical medicine: Waterloo participation in CLEF eHealth 2017. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Cotik V, Vivaldi J, Rodriguez H (2015) Semantic tagging of French medical entities using distant learning. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- De Vine L, Kholghi M, Zuccon G, Sitbon L, Nguyen A (2015) Patient empowerment via technologies for patient-friendly personalized language. In: Proceedings of the 13th annual workshop of the Australasian Language Technology Association. Australasian Language Technology Association, Sydney, pp 153–164
- Demner-Fushman D, Elhadad N (2016) Aspiring to unintended consequences of natural language processing: a review of recent developments in clinical and consumer-generated text processing. *Yearb Med Inform* (1):224–233. <https://doi.org/10.15265/IY-2016-017>
- Dermouche M, Looten V, Flicoteaux R, Chevret S, Velcin J, Taright N (2016) ECSTRA-INSERM @ CLEF eHealth2016-task 2: ICD10 code extraction from death certificates. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- D'Hondt E, Grau B, Zweigenbaum P (2015a) LIMSI @ CLEF eHealth 2015 — task 2. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- D'Hondt E, Morlane-Hondère F, Campillos L, Bouamor D, Ribeiro S, Lavergne T (2015b) LIMSI @ CLEF eHealth 2015 — task 1b. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Di Nunzio G, Beghini F, Vezzani F, Henrot G (2017a) An interactive two-dimensional approach to query aspects rewriting in systematic reviews. IMS Unipd at CLEF eHealth task 2. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Di Nunzio G, Beghini F, Vezzani F, Henrot G (2017b) A lexicon based approach to classification of ICD10 codes. IMS Unipd at CLEF eHealth task 1. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Diaz-Galiano M, Martin-Valdivia M, Jiménez-Zafra S, Andreu A, Urena-López L (2017) SINAI at CLEF eHealth 2017 task 3s. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Dramé K, Mougin F, Diallo G (2014) Query expansion using external resources for improving information retrieval in the biomedical domain. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Ebersbach M, Herms R, Lohr C, Eibl M (2016) Wrappers for feature subset selection in CRF-based clinical information extraction. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Ebersbach M, Herms R, Eibl M (2017) Fusion methods for ICD10 code classification of death certificates in multilingual corpora. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Fan J, Sood N, Huang Y (2013) Disorder concept identification from clinical notes: an experience with the ShARE/CLEF 2013 challenge. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Filannino M, Uzuner Ö (2018) Advancing the state of the art in clinical natural language processing through shared tasks. *Yearb Med Inform* 27(01):184–192. <https://doi.org/10.1055/s-0038-1667079>
- Friberg Heppin K, Järvelin A (2012) Towards improving search results for medical experts and laypersons. In: CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Ghoddousi A, Huang J (2015) York University at CLEF 2015 eHealth: medical document retrieval. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>

- Goeriot L, Jones G, Kelly L, Leveling J, Hanbury A, Müller H, Salanterä S, Suominen H, Zuccon G (2013a) ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: information retrieval to address patients' questions when reading clinical reports. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Goeriot L, Kelly L, Jones G, Zuccon G, Suominen H, Hanbury A, Müller H, Leveling J (2013b) Creation of a new evaluation benchmark for information retrieval targeting patient information needs. In: Proceedings of the 5th international workshop on evaluating information access (EVIa), A satellite workshop of the NTCIR-10 conference. National Institute of Informatics/Kijima Printing, Tokyo/Fukuoka
- Goeriot L, Kelly L, Leveling J (2014a) An analysis of query difficulty for information retrieval in the medical domain. In: SIGIR'14: proceedings of the 37th international ACM SIGIR conference on research and development in information retrieval. Association for Computer Machinery, New York, pp 1007–1010
- Goeriot L, Kelly L, Li W, Palotti J, Pecina P, Zuccon G, Hanbury A, Jones G, Müller H (2014b) ShARe/CLEF eHealth evaluation lab 2014, task 3: user-centred health information retrieval. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Goeriot L, Kelly L, Suominen H, Hanlen L, Névéol A, Grouin C, Palotti JRM, Zuccon G (2015) Overview of the CLEF eHealth Evaluation Lab 2015. In: Mothe J, Savoy J, Kamps J, Pinel-Sauvagnat K, Jones GJF, SanJuan E, Cappellato L, Ferro N (eds) (2015) Experimental IR meets multilinguality, multimodality, and interaction. In: Proceedings of the sixth international conference of the CLEF association (CLEF 2015). Lecture notes in computer science (LNCS), vol 9283. Springer, Heidelberg, pp 429–443
- Goeriot L, Kelly L, Zuccon G, Palotti J (2016) Building evaluation datasets for consumer-oriented information retrieval. In: Proceedings of the tenth international conference on language resources and evaluation (LREC 2016), European Language Resources Association (ELRA), Paris
- Goeriot L, Kelly L, Suominen H, Névéol A, Robert A, Kanoulas E, Spijker R, Palotti JRM, Zuccon G (2017) CLEF 2017 eHealth evaluation lab overview. In: Jones GJF, Lawless S, Gonzalo J, Kelly L, Goeriot L, Mandl T, Cappellato L, Ferro N (eds) (2017) Experimental IR meets multilinguality, multimodality, and interaction. In: Proceedings of the eighth international conference of the CLEF association (CLEF 2017). Lecture notes in computer science (LNCS), vol 10456. Springer, Heidelberg, pp 291–303
- Gung J (2013) Using relations for identification and normalization of disorders: team CLEAR in the ShARe/CLEF 2013 eHealth evaluation lab. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Hamon T, Grouin C, Zweigenbaum P (2014) Disease and disorder template filling using rule-based and statistical approaches. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Hanbury A, Müller H (2012) Khresmoi — multimodal multilingual medical information search. In: MIE village of the future
- Hanlen L (2012) National eHealth living lab: removing the “e” from e-health. In: CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Heimonen J, Salakoski T, Salanterä S (2012) An ontology to improve accessibility and quality of patient instructions. In: CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Herbst K, Fähnrich C, Neves M, Schapranow M (2014) Applying in-memory technology for automatic template filling in the clinical domain. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Hermes R, Richter D, Eibl M, Ritter M (2015) Search for clinical speech recognition. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>

- Hervas L, Martinez V, Sanchez I, Diaz A (2013a) UCM at CLEF eHealth 2013 shared task1a. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Hervas L, Martinez V, Sanchez I, Diaz A (2013b) UCM at CLEF eHealth 2013 shared task1b. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Ho-Dac L, Tanguy L, Grauby C, Mby N, Malosse J, Rivière L, Veltz-Mauclair A, Wauquier M (2016) LITL at CLEF eHealth2016: recognizing entities in French biomedical documents. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Ho-Dac L, Fabre C, Birski A, Boudraa I, Bourriot A, Cassier M, Delvenne L, Garcia-Gonzalez C, Kang E, Piccinini E, Rohrbacher C, S'egui A (2017) LITL at CLEF eHealth2017: automatic classification of death reports. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Hollmann N, Eickhoff C (2017) Ranking and feedback-based stopping for recall-centric document retrieval. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Huang CC, Lu Z (2016) Community challenges in biomedical text mining over 10 years: success, failure and the future. *Brief Bioinform* 17(1):132–144
- Huynh H, Ho S (2014) ShARe/CLEFeHealth: a hybrid approach for task 2. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Huynh N, Nguyen T, Ho Q (2015) TeamHCMUS: a concept-based information retrieval approach for web medical documents. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Hyman H, Fridy W (2014) An eHealth process model of visualization and exploration to support improved patient discharge record understanding and medical knowledge enhancement. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Ilic D (2010) The role of the internet on patient knowledge management, education, and decision-making. *Telemed J E Health* 16(6):664–669
- Isenius N, Kvist M, Velupillai S (2012) Initial results in the development of SCAN: a Swedish clinical abbreviation normalizer. In: CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Jagannathan V, Ganesan D, Jagannathan A, Kavi R, Lamb A, Peters F, Seeger S (2013) WVU NLP class participation in ShARe/CLEF challenge. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Jain D (2015) Supervised named entity recognition for clinical data. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Jiang J, YGuan, Zhao C (2015) WI-ENRE in CLEF eHealth evaluation lab 2015: clinical named entity recognition based on CRF. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Jimmy, Zuccon G, Koopman B (2017) QUT ielab at CLEF 2017 e-Health IR task: knowledge base retrieval for consumer health search. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Johri N, Niwa Y, Raghavendra Chikka V (2014) Optimizing Apache cTAKES for disease/disorder template filling: team HITACHI in the ShARe/CLEF 2014 eHealth evaluation lab. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Jones G, Karlgren J, Kreiner K, Müller H, Salanterä S (2012) Panel presentation: towards systematic evaluation of methods, applications, and resources for eHealth document analysis. In: CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>

- Jonnagaddala J, Hu F (2017) Automatic coding of death certificates to ICD-10 terminology. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Kalpathy-Cramer J, Müller H, Bedrick S, Eggel I, Garcia Seco de Herrera A, Tsirikika T (2011) Overview of the CLEF 2011 medical image classification and retrieval tasks. In: Petras V, Forner P, Clough P, Ferro N (eds) CLEF 2011 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1177/>
- Kanhov M, Feng X, Dalianis H (2012) Natural language generation from SNOMED specifications. In: CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Kanoulas E, Li D, Azzopardi L, Spijker R (2017) CLEF 2017 technologically assisted reviews in empirical medicine overview. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Kelly L, Goeuriot L, Jones G, Hanbury A (2012) Considering subjects and scenarios in large-scale user-centered evaluation of a multilingual multimodal medical search system. In: CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Kelly L, Goeuriot L, Suominen H, Schreck T, Leroy G, Mowery DL, Velupillai S, Webber Chapman W, Martínez D, Zuccon G, Palotti JRM (2014) Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. In: Kanoulas E, Lupu M, Clough P, Sanderson M, Hall M, Hanbury A, Toms E (eds) (2014) Information access evaluation – multilinguality, multimodality, and interaction. In: Proceedings of the fifth international conference of the CLEF initiative (CLEF 2014). Lecture notes in computer science (LNCS), vol 8685. Springer, Heidelberg, pp 172–191
- Kelly L, Goeuriot L, Suominen H, Névéol A, Palotti J, Zuccon G (2016) Overview of the CLEF eHealth Evaluation Lab 2016. In: Fuhr N, Quresma P, Gonçalves T, Larsen B, Balog K, Macdonald C, Cappellato L, Ferro N (eds) (2016) Experimental IR meets multilinguality, multimodality, and interaction. In: Proceedings of the seventh international conference of the CLEF association (CLEF 2016). Lecture notes in computer science (LNCS), vol 9822. Springer, Heidelberg, pp 255–266
- Kholghi M, Sitbon L, Zuccon G, Nguyen A (2014) Factors influencing robustness and effectiveness of conditional random fields in active learning frameworks. In: AusDM: the twelfth Australasian data mining conference, Queensland University of Technology, Brisbane, pp 1007–1010
- Kholghi M, Sitbon L, Zuccon G, Nguyen A (2015) External knowledge and query strategies in active learning: a study in clinical information extraction. In: Proceedings of the 24th ACM international on conference on information and knowledge management. Association for Computing Machinery, New York, pp 143–152
- Kholghi M, Sitbon L, Zuccon G, Nguyen A (2016) Active learning: a step towards automating medical concept extraction. *J Am Med Inform Assoc* 23(2):289–296
- Koopman B, Zuccon G (2014) Relevation!: An open source system for information retrieval relevance assessment. In: SIGIR'14: proceedings of the 37th international ACM SIGIR conference on research and development in information retrieval. Association for Computer Machinery, New York, pp 1243–1244
- Kreiner K, Eckmann H, Hayn D, Kastner P (2012) On the use of text messaging in a diabetes telehealth system results and evaluation of a content analysis. In: CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Ksentini N, Tmar M, Boughanem M, Gargouri F (2015) Miracl at CLEF 2015: user-centred health information retrieval task. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Laippala V, Danielsson-Ojala R, Lundgrén-Laine H, Salanterä S, Salakoski T (2012) Vocabulary in discharge documents: the patient's perspective. In: CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>

- Leaman R, Khare R, Lu Z (2013) NCBI at 2013 ShARe/CLEF eHealth shared task: disorder normalization in clinical notes with dnorm. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Lee G (2017) A study of convolutional neural networks for clinical document classification in systematic reviews: SysReview at CLEF eHealth 2017. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Limsopatham N, Macdonald C, Ounis I (2013) University of Glasgow at CLEF 2013: experiments in eHealth task 3 with Terrier. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Liu N, Ku L (2014) CLEFeHealth 2014 normalization of information extraction challenge using multi-model method. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Liu X, Nie J (2015) Bridging layperson's queries with medical concepts — GRIUM@CLEF2015 eHealth task 2. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Liu H, Waghlikar K, Jonnalagadda S, Sohn S (2013) Integrated cTAKES for concept mention detection and normalization. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Liu X, Liu X, Shen W, Nie J (2014) Mining disorder attributes with rules and statistical learning. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Lu F (2015) Employing query expansion models to help patients diagnose themselves. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Luu T, Phan R, Davey R, Chetty G (2015) Automatic clinical speech recognition for CLEF 2015 eHealth challenge. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Malagon J, L'opez M (2014) LABERINTO at ShARe/CLEF eHealth evaluation lab 2014. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Martinez D, Suominen H, Ananda-Rajah M, Cavedon L (2012) Biosurveillance for invasive fungal infections via text mining. In: CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- McAllister M, Dunn G, Payne K, Davies L, Todd C (2012) Patient empowerment: the need to consider it as a measurable patient-reported outcome for chronic conditions. *BMC Health Serv Res* 12:157
- Miftahutdinov Z, Tutubalina E (2017) KFU at CLEF eHealth 2017 task 1: ICD-10 coding of English death certificates with recurrent neural networks. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Mkrtchyan T, Sonntag D (2014) Deep parsing at the CLEF2014 IE task. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Moen H, Marsi E (2012) Towards retrieving and ranking clinical recommendations with cross-lingual random indexing. In: CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Mottin L, Gobeill J, Mottaz A, Pasche E, Gaudinat A, Ruch P (2016) BiTeM at CLEF eHealth evaluation lab 2016 task 2: multilingual information extraction. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Mowery D, South B, Christensen L, Murtola L, Salanterä S, Suominen H, Martinez D, Elhadad N, Pradhan S, Savova G, Chapman W (2013) Task 2: ShARe/CLEF eHealth evaluation lab 2013. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>

- Mowery D, Velupillai S, South B, Christensen L, Martinez D, Kelly L, Goeriot L, Elhadad N, Pradhan S, Savova G, Chapman W (2014) Task 2: ShARe/CLEF eHealth evaluation lab 2014. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Mowery DL, South BR, Christensen L, Leng J, Peltonen LM, Salanterä S, Suominen H, Martinez D, Velupillai S, Elhadad N, Savova G, Pradhan S, Chapman WW (2016) Normalizing acronyms and abbreviations to aid patient understanding of clinical texts: ShARe/CLEF eHealth challenge 2013, task 2. *J Biomed Semant* 7(1):43
- Murtola L, Kauhanen L, Moen H, Lundgr'en-Laine H, Salakoski T, Salanterä S (2013) Using text mining to explore factors associated with acute confusion in cardiac patients documentation. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Nesrine K, Mohamed T, Faiez G (2014) Miracl at CLEF 2014: eHealth information retrieval task. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Névóel A, Grouin C, Leixa J, Rosset S, Zweigenbaum P (2014) The QUAERO French medical corpus: a resource for medical entity recognition and normalization. In: *Proceeding of BioTextMining Work*, pp 24–30
- Névóel A, Grouin C, Tannier X, Hamon T, Kelly L, Goeriot L, Zweigenbaum P (2015) CLEF eHealth evaluation lab 2015 task 1b: clinical named entity recognition. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Névóel A, Cohen K, Grouin C, Hamon T, Lavergne T, Kelly L, Goeriot L, Rey G, Robert A, Tannier X, Zweigenbaum P (2016) Clinical information extraction at the CLEF eHealth evaluation lab 2016. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Névóel A, Robert A, Anderson R, Cohen K, Grouin C, Lavergne T, Rey G, Rondet C, Zweigenbaum P (2017) CLEF eHealth 2017 multilingual information extraction task overview: ICD10 coding of death certificates in English and French. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Norman C, Leeflang M, N'evóel A (2017) LIMSI@CLEF eHealth 2017 task 2: logistic regression for automatic article ranking. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Oh H, Jung Y (2014) Ta multiple-stage approach to re-ranking clinical documents. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Oh H, Jung Y (2016) KISTI at CLEF eHealth 2016 task 3: ranking medical documents using word vectors. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Oh H, Jung Y (2017) KISTI at CLEF eHealth 2017 patient-centered information retrieval task-1: improving medical document retrieval with query expansion. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Oh H, Jung Y, Kim K (2015) KISTI at CLEF eHealth 2015 task 2. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Osborne J (2014) Disease template filling using the CTAKES YTEX Branch for ShareCLEF 2014 task 2a. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Osborne J, Gyawali B, Solorio T (2013) Evaluation of YTEX and MetaMap for clinical concept recognition. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>

- Ozturkmenoglu O, Alpkocak A, Kilinc D (2014) DEMIR at CLEF eHealth: the effects of selective query expansion to information retrieval. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Palotti J, Rekabsaz N (2017) Exploring understandability features to personalize consumer health search. TUW at CLEF 2017 eHealth. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Palotti J, Zuccon G, Goeuriot L, Kelly L, Hanbury A, Jones G, Lupu M, Pecina P (2015a) CLEF eHealth evaluation lab 2015, task 2: retrieving information about medical symptoms. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Palotti J, Zuccon G, Hanbury A (2015b) The influence of pre-processing on the estimation of readability of web documents. In: Proceedings of the 24th ACM international on conference on information and knowledge management. Association for Computing Machinery, New York, pp 1763–1766
- Palotti J, Goeuriot L, Zuccon G, Hanbury A (2016a) Patient empowerment via technologies for patient-friendly personalized language. In: Proceedings of the 39th international ACM SIGIR conference. Association for Computing Machinery, New York, pp 965–968
- Palotti J, Zuccon G, Bernhardt J, Hanbury A, Goeuriot L (2016b) Assessors agreement: a case study across assessor type, payment levels, query variations and relevance dimensions. In: Fuhr N, Quaresma P, Gonçalves T, Larsen B, Balog K, Macdonald C, Cappellato L, Ferro N (eds) (2016) Experimental IR meets multilinguality, multimodality, and interaction. In: Proceedings of the seventh international conference of the CLEF association (CLEF 2016). Lecture notes in computer science (LNCS), vol 9822. Springer, Heidelberg, pp 40–53
- Palotti J, Zuccon G, Jimmy, Pecina P, Lupu M, Goeuriot L, Kelly L, Hanbury A (2017) CLEF 2017 task overview: the IR task at the eHealth evaluation lab — evaluating retrieval methods for consumer health search. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Patrick J, Safari L, Ou Y (2013a) ShARe/CLEF eHealth 2013 named entity recognition and normalization of disorders challenge. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Patrick J, Safari L, Ou Y (2013b) Share/clef ehealth 2013 normalization of acronyms/abbreviations challenge. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Pestian J, Matykiewicz P, Linn-Gust M, South B, Uzuner O, Wiebe J, Cohen K, Hurdle J, Brew C (2011) Sentiment analysis of suicide notes: a shared task. *Biomed Inform Insights* 5(Suppl. 1):3–16
- Pradhan S, Elhadad N, South B, Martinez D, Christensen L, Vogel A, Suominen H, Chapman W, Savova G (2013) Task 1: ShARe/CLEF eHealth evaluation lab 2013. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Pradhan S, Elhadad N, South BR, Martinez D, Christensen L, Vogel A, Suominen H, Chapman WW, Savova G (2015) Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *J Am Med Inform Assoc* 22(1):143–154. <https://doi.org/10.1136/amiajnl-2013-002544>
- Quiroz L, Mennes L, Dehghani M, Kanoulas E (2016) Distributional semantics for medical information extraction. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Ramanan S, Senthil Nathan P (2014) Cocoa: Extending a rule-based system to tag disease attributes in clinical records. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Ramanan S, Broido S, Senthil Nathan P (2013) Performance of a multi-class biomedical tagger on clinical records. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>



- Rekabsaz N, Lupu M, Hanbury A, Zuccon G (2016) Generalizing translation models in the probabilistic relevance framework. In: Proceedings of the 25th ACM international on conference on information and knowledge management. Association for Computing Machinery, New York, pp 711–720
- Roberts PM, Cohen AM, Hersh WR (2009) Tasks, topics and relevance judging for the TREC genomics track: five years of experience evaluating biomedical text information retrieval systems. *Inf Retr* 12:81–97
- Robertson S, Hull D (2000) The TREC-9 filtering track final report. In: NIST Special Publication 500-249: The 9th Text REtrieval Conference (TREC 9). National Institute of Standards and Technology, Gaithersburg, pp 25–40
- Saleh S, Pecina P (2014) CUNI at the ShARe/CLEF eHealth evaluation lab 2014. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Saleh S, Pecina P (2016) Reranking hypotheses of machine-translated queries for cross-lingual information retrieval. In: Fuhr N, Quaresma P, Gonçalves T, Larsen B, Balog K, Macdonald C, Cappellato L, Ferro N (eds) (2016) Experimental IR meets multilinguality, multimodality, and interaction. In: Proceedings of the seventh international conference of the CLEF association (CLEF 2016). Lecture notes in computer science (LNCS), vol 9822. Springer, Heidelberg, pp 54–68
- Saleh S, Pecina P (2017) Task3 patient-centred information retrieval: team CUNI. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Saleh S, Bibyna F, Pecina P (2015) CUNI at the CLEF 2015 eHealth lab task 2. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Scells H, Zuccon G, Deacon A, Koopman B (2017) QUT ielab at CLEF 2017 technology assisted reviews track: initial experiments with learning to rank. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Sequeira J, Miranda N, Goncalves T, Quaresma P (2014) Team UEvora at CLEF eHealth 2014 task2a. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Seva J, Kittner M, Roller R, Leser U (2017) Multi-lingual ICD-10 coding using a hybrid rule-based and supervised classification approach at CLEF eHealth 2017. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Shenwei W, Nie J, Liu X, Liu X (2014) An investigation of the effectiveness of concept-based approach in medical information retrieval. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Silva R, Lopes C (2016) The effectiveness of query expansion when searching for health related content: InfoLab at CLEF eHealth 2016. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Singh J, Thomas L (2017) IIIT-H at CLEF eHealth 2017 task 2: technologically assisted reviews in empirical medicine. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Singh G, Marshall I, Thomas J, Wallace B (2017) Identifying diagnostic test accuracy publications using a deep model. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Soldaini L, Edman W, Goharian N (2016) Team GU-IRLAB at CLEF eHealth 2016: task 3. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Song Y, He Y, Hu Q, He L, Haacke E (2015) ECNU at 2015 eHealth task 2: user-centred health information retrieval. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>

- Song Y, He Y, Liu H, Hu Q, He L, Wang Y (2016a) ECNU at 2016 eHealth task 1: handover information extraction. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Song Y, He Y, Liu H, Wang Y, Hu Q, He L (2016b) ECNU at 2016 eHealth task 3: patient-centred information retrieval. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Soualmia L, Cabot C, Dahamna B, Darmoni S (2015) SIBM at CLEF e-Health evaluation lab 2015. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- South BR, Mowery D, Suo Y, Leng J, Óscar Ferrández, Meystre SM, Chapman WW (2014) Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text. *J Biomed Inform* 50:162–172. <https://doi.org/10.1016/j.jbi.2014.05.002>, special issue on Informatics Methods in Medical Privacy
- Sun W, Rumshisky A, Uzuner O (2013) Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *J Am Med Inform Assoc* 20(5):806–813. <https://doi.org/10.1136/amiajnl-2013-001628>
- Suominen H (2012a) CLEFeHealth2012 - the CLEF 2012 workshop on cross-language evaluation of methods, applications, and resources for eHealth document analysis. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) (2012) CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Suominen H (2012b) Towards international privacy-preserving benchmarks of eHealth technologies. In: CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Suominen H (2014) Text mining and information analysis of health documents. *Artif Intell Med* 61(3):127–130
- Suominen H, Basilakis J, Johnson M, Dawson L, Hanlen L, Kelly B, Yeo A, Sanchez P (2012a) Clinical speech to text evaluation setting. In: CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Suominen H, Kreiner K, Hanlen L (2012b) Towards ease of building legos in assessing eHealth language technologies: a RESTful laboratory for data and software. In: CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Suominen H, Salanterä S, Velupillai S, Webber Chapman W, Savova GK, Elhadad N, Pradhan S, South BR, Mowery DL, Jones GJF, Leveling J, Kelly L, Goeuriot L, Martínez D, Zuccon G (2013) Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: Forner P, Müller H, Paredes R, Rosso P, Stein B (eds) (2013) Information access evaluation meets multilinguality, multimodality, and visualization. In: Proceedings of the fourth international conference of the CLEF initiative (CLEF 2013). Lecture notes in computer science (LNCS), vol 8138. Springer, Heidelberg, pp 212–231
- Suominen H, Schreck T, Leroy G, Hochheiser H, Goeuriot L, Kelly L, Mowery D, Nualart J, Ferraro G, Keim D (2014) Task 1 of the ShARe/CLEF eHealth evaluation lab 2014. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Suominen H, Hanlen L, Goeuriot L, Kelly L, Jones G (2015a) Task 1a of the CLEF eHealth evaluation lab 2015: clinical speech recognition. In: CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Suominen H, Zhou L, Hanlen L, Ferraro G (2015b) Benchmarking clinical speech recognition and information extraction: new data, methods, and evaluations. *JMIR Med Inform* 3(2):e19
- Suominen H, Zhou L, Goeuriot L, Kelly L (2016) Task 1 of the CLEF eHealth evaluation lab 2016: handover information extraction. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Suominen H, Kelly L, Goeuriot L (2018a) Scholarly influence of the Conference and Labs of the Evaluation Forum eHealth initiative: review and bibliometric study of the 2012 to 2017 outcomes. *JMIR Res Protoc* 7(7):e10961. <https://doi.org/10.2196/10961>

- Suominen H, Kelly L, Goeriot L, Névéol A, Ramadier L, Robert A, Kanoulas E, Spijker R, Azzopardi L, Li D, Jimmy, Palotti J, Zucon G (2018b) Overview of the CLEF eHealth Evaluation Lab 2018. In: Bellot P, Trabelsi C, Mothe J, Murtagh F, Nie JY, Soulier L, SanJuan E, Cappellato L, Ferro N (eds) *Experimental IR meets multilinguality, multimodality, and interaction. Proceedings of the ninth international conference of the CLEF association (CLEF 2018)*, Lecture notes in computer science (LNCS), vol 11018. Springer, Heidelberg, pp 286–301
- Tang B, Wu Y, Jiang M, Denny J, Xu H (2013) Recognizing and encoding disorder concepts in clinical text using machine learning and vector space model. In: *CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org)*, ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Tchechmedjiev A, Abdaoui A, Emonet V, Jonquet C (2017) ICD10 coding of death certificates with the NCBO and SIFR annotator(s) at CLEF eHealth 2017 task 1. In: *CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org)*, ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Thakkar H, Iyer G, Shah K, Majumder P (2014) Team IRLabDAICT at ShARe/CLEF eHealth 2014 task 3: user-centered information retrieval system for clinical documents. In: *CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org)*, ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Thesprasith O, Jaruskulchai C (2014) CSKU GPRF-QE for medical topic web retrieval. In: *CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org)*, ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Thesprasith O, Jaruskulchai C (2015) Task 2a: Team KU-CS: query coherence analysis for PRF and genomics expansion. In: *CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org)*, ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Thuma E, Anderson G, Mosweunyane G (2015) UBML participation to CLEF eHealth IR challenge 2015: task 2. In: *CLEF 2015 working notes, CEUR workshop proceedings (CEUR-WS.org)*, ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Thuma E, Motlogelwa N, Leburu-Dingalo T (2016) Task 3: patient-centered information retrieval, IRTask 1: ad-hoc search — TEAM ub-botswana. In: *CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org)*, ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Thuma E, Motlogelwa N, Leburu-Dingalo T (2017) Ub-Botswana participation to clef ehealth IR challenge 2017: task 3 (irtask1: Ad-hoc search). In: *CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org)*, ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Tsikrika T, Larsen B, Müller H, Endrullis S, Rahm E (2013) The scholarly impact of CLEF (2000–2009). In: Forner P, Müller H, Paredes R, Rosso P, Stein B (eds) (2013) *Information access evaluation meets multilinguality, multimodality, and visualization. In: Proceedings of the fourth international conference of the CLEF initiative (CLEF 2013). Lecture notes in computer science (LNCS)*, vol 8138. Springer, Heidelberg, pp 1–12
- Ullah M, Aono M (2016) KDEIR at CLEF eHealth 2016: health documents re-ranking based on query variations. In: *CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org)*, ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Uzuner O, South BR, Shen S, DuVall S (2011) 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 18(5):552–556
- van Altena A, Delgado Olabarriga S (2017) Predicting publication inclusion for diagnostic accuracy test reviews using random forests and topic modelling. In: *CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org)*, ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- van Mulligen E, Afzal Z, Akhondi S, Vo D, Kors J (2016) Erasmus MC at CLEF eHealth 2016: concept recognition and coding in French texts. In: *CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org)*, ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Verberne S (2014) A language-modelling approach to user-centred health information retrieval. In: *CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org)*, ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>

- Vivaldi J, Rodriguez H, Cotik V (2016) Semantic tagging and normalization of French medical entities. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Voorhees E, Tong RM (2011) Overview of the TREC 2011 medical records track. In: NIST Special Publication 500-296: the text Retrieval conference 2011. National Institute of Standards and Technology, Gaithersburg
- Wang C, Akella R (2013) UCSC's system for CLEF eHealth 2013 task 1. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Wang R, Lu W, Ren K (2016a) WHUIRGroup at the CLEF 2016 eHealth lab task 3. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Wang Y, Wu S, Liu H (2016b) MayoNLPTeam at the 2016 CLEF eHealth information retrieval task 1. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Wu J, Huang J (2014) Task 3a: Team YORKU. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Wu Y, Tang B, Jiang M, Moon S, Denny J, Xu H (2013) Clinical acronym/abbreviation normalization using a hybrid approach. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Xia Y, Zhong X, Liu P, Tan C, Na S, Hu Q, Huang Y (2013a) Combining MetaMap and cTAKES in disorder recognition: THCIB at CLEF eHealth lab 2013 task 1. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Xia Y, Zhong X, Liu P, Tan C, Na S, Hu Q, Huang Y (2013b) Normalization of abbreviations/acronyms: THCIB at CLEF eHealth 2013 task 2. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Yang H, Goncalves T (2017) UEvora at CLEF eHealth 2017 task 3. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Yang C, Bhattacharya S, Srinivasan P (2014) The university of Iowa at CLEF 2014: eHealth task 3. In: CLEF 2014 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>
- Yu Z, Menzies T (2017) Data balancing for technologically assisted reviews: undersampling or reweighting. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Zhang G, Ding Y, Milojević S (2013a) Citation content analysis (CCA): a framework for syntactic and semantic analysis of citation content. *J Am Soc Inf Sci Technol* 64(7):1490–1503
- Zhang Y, Cohen T, Jiang M, Tang B, Xu H (2013b) Evaluation of vector space models for medical disorders information retrieval. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Zhong X, Xia Y, Xie Z, Na S, Hu Q, Huang Y (2013) Concept-based medical document retrieval: THCIB at CLEF eHealth lab 2013 task 3. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Zhou L, Suominen H (2015) Information extraction to improve standard compliance: the case of clinical handover. In: Proceedings of the 28th Australasian joint conference on artificial intelligence 2015 (AI2015). Lecture notes in computer science, vol 9457. Springer, Heidelberg, pp 644–649
- Zhou L, Suominen H, Hanlen L (2015) Evaluation data and benchmarks for cascaded speech recognition and entity extraction. In: Proceedings of the ACM multimedia 2015 workshop on speech, language and audio in multimedia, vol 10. Association for Computing Machinery, New York, pp 15–18

- Zhu D, Wu S, Masanz J, Carterette B, Liu H (2013) Using discharge summaries to improve information retrieval in clinical domain. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Zuccon G (2016) Understandability biased evaluation for information retrieval. In: European conference on information retrieval (ECIR 2016). Lecture notes in computer science, vol 9626. Springer, Heidelberg, pp 280–292
- Zuccon G, Koopman B (2014) Integrating understandability in the evaluation of consumer health search engines. In: Medical information retrieval (MedIR) workshop. Association for Computer Machinery, New York
- Zuccon G, Holloway A, Koopman B, Nguyen A (2013a) Identify disorders in health records using conditional random fields and Metamap AEHRC at ShARe/CLEF 2013 eHealth evaluation lab task 1. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Zuccon G, Koopman B, Nguyen A (2013b) Retrieval of health advice on the web AEHRC at ShARe/CLEF eHealth evaluation lab task 3. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Zuccon G, Koopman B, Palotti J (2015) Diagnose this if you can: on the effectiveness of search engines in finding medical self-diagnosis information. In: European conference on information retrieval (ECIR 2015). Lecture notes in computer science, vol 9022. Springer, Heidelberg, pp 562–567
- Zuccon G, Palotti J, Goeriot L, Kelly L, Lupu M, Pecina P, Müller H, Budaher J, Deacon A (2016a) The IR task at the CLEF eHealth evaluation lab 2016: User-centred health information retrieval. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Zuccon G, Palotti J, Hanbury A (2016b) Query variations and their effect on comparing information retrieval systems. In: Proceedings of the 25th ACM international on conference on information and knowledge management. Association for Computing Machinery, New York, pp 691–700
- Zweigenbaum P, Lavergne T (2016) LIMSI ICD10 coding experiments on C/epiDC death certificate statements. In: CLEF 2016 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1609/>
- Zweigenbaum P, Lavergne T (2017) Multiple methods for multi-class, multi-label ICD-10 coding of multi-granularity, multilingual death certificates. In: CLEF 2017 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>
- Zweigenbaum P, Deleger L, Lavergne T, Neveol A, Bodnari A (2013) A supervised abbreviation resolution system for medical text. In: CLEF 2013 working notes, CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>

# Multilingual Patent Text Retrieval Evaluation: CLEF-IP



Florina Piroi and Allan Hanbury

**Abstract** The CLEF-IP evaluation lab ran between 2009 and 2013 with a two-fold expressed purpose: (a) to encourage research in the area of patent retrieval with a focus on cross language retrieval, and (b) to provide a large and clean data set of patent related data, in the three main European languages, for experimentation. In its first year, CLEF-IP organized one task only, a text retrieval task that modelled the “Search for Prior Art” done by experts at patent offices. In the following years the types of CLEF-IP tasks broadened to include patent text classification, patent image retrieval and classification, and (formal) structure recognition. With each task, the test collection was extended to accommodate for the additional tasks. In this chapter we overview the evaluation tasks dealing with the textual content of the patents. The Intellectual Property (IP) domain is one where specific expertise is critical, implementing Information Retrieval (IR) approaches to support some of its tasks cannot be done without the use of this domain know-how. Even when such know-how is at hand, retrieval results, in general, do not come close to the expectations of patent experts.

## 1 Introduction

In a nutshell, patents can be seen as contracts between inventors and governments by which the former can exclude other parties from manufacturing and exploiting an invention without permission from the patent owner. This corresponds to a pessimistic view of the patent system based on a “blocking effect,” raising a sequence of issues in the modern world, like, for example, invention fragmentation or failures in securing patent licensing (Galasso and Schankerman 2013). On the more optimistic note, the patent system is viewed as fundamental to the diffusion

---

F. Piroi (✉) · A. Hanbury  
Institute of Information Systems Engineering, TU Wien, Vienna, Austria  
e-mail: [florina.piroi@tuwien.ac.at](mailto:florina.piroi@tuwien.ac.at); [allan.hanbury@tuwien.ac.at](mailto:allan.hanbury@tuwien.ac.at)

of ideas and a key incentive to advancing the technological knowledge of society, which some countries underline in their patent law (Kumagai 2005).

In this chapter we look, first, at how the patent system evolved to its current state (Sect. 2.1), then we look at search and retrieval on patent data (Sect. 2.2). In Sect. 2.3 we outline the main phases of the patent's life-cycle and recount the administrative character of the patent systems. We continue with describing the CLEF-IP test collection and describe the text retrieval tasks that were organized in this evaluation lab in Sects. 3 and 4. We finish with a description of the submissions and a submission scores summary.

## 2 A Background on Patents

As a main governmental instrument to increase research and development (Galasso and Schankerman 2013), patents are not only an output of R&D activities but also an indicator of the technological competitiveness at national, regional or sectoral levels (Frietsch et al. 2010). In this section we give an abridged account of the origins of the modern patent systems. We, then, explain the need for Information Retrieval research in the patent domain, giving an account on the IR research efforts in the IP domain. The section continues with a description of patent data characteristics.

### 2.1 *The Patent System: A Very Brief History*

Inventions, as the root of new technologies and developments, provide consistent input to civilization advancements. Until the emergence of the Greek civilization, discoveries and inventive activities were extremely low paced (Skolnik 1977). The first recorded grant of a monopoly refers to the time of the Sybarites (approximately 750 B.C. (Pfaller 2013b) and (Anthon 1841)) when 1 year exclusiveness on exceptional food recipes were awarded.

The emergence of the Greek civilization accelerated the pace of discovery, but the idea of invention was established only by the end of the thirteenth century, at the beginning of the Renaissance (Skolnik 1977). Historians agree that one of the first exclusive rights of use we know about was awarded to a Florentine architect, Filippo Brunelleschi, for a special type of barge that was capable of transporting heavy loads (marble) along the Arno River (Skolnik 1977; Pfaller 2013a). In the exposure of motives to grant this patent it was shown that the inventor was refusing to reveal his invention for fear that there was not enough protection against others who would replicate and use it. The period of exclusiveness awarded to Brunelleschi was of 3 years.

A few decades later, the first patent system was developed in fifteenth century Venice and was explicitly utilised to promote innovation (May 2010). In March 1474, the Senate of Venice issued a decree which made patents a subject of a generalized law instead of individual petitions and monopoly grants (May 2010; Skolnik 1977).

Other patent laws (in the sense we give it today) were the “Statute of Monopolies,” released in 1623, in England, and in 1787, in France, which granted longer periods of exclusive use for inventions (Rich 1993; Skolnik 1977). In America, where the first patent was granted as early as 1641, the first US Patent Act was passed in April 1790, and conferred inventors exclusive rights for 14 years for disclosing their inventions (Skolnik 1977; Mossoff 2007). Later, in 1861, this time period was extended to 17 years. Other European countries also extended and modernized their patent and monopolies laws during the nineteenth century, and during the twentieth century, the use of the patent system became worldwide ubiquitous (Hall 2017).

As national patent systems evolved, the differences in patent laws between nations were considerable. The 1883 Paris Convention for the Protection of Industrial Property successfully established a unified system for multinational filings, enabling worldwide priority to be obtained for an invention originating in any one country part of the treaty (Skolnik 1977; Hall 2017). A 110 years later, in 1995, the adoption of the Trade-Related Aspects of Intellectual Property Rights Agreement (TRIPS), ensures that the patent granting process is approximately the same everywhere in the world (Hall 2017). The World Intellectual Property Organization (WIPO) joined the administrative offices of the Paris Convention of 1883 and the Berne Convention of 1886, which established rules about protection of literary and artistic works. WIPO’s<sup>1</sup> first international IP filing service was launched with the adoption of the Madrid Agreement in 1891. In 1978 the Patent Cooperation Treaty (PCT) came into existence, which allows inventors in any of the treaty’s signatory countries to file patent applications and seek protection of the invention in countries other than the country of origin (PCT 1970).

Patent rights, when granted, are usually restricted within the border of the patent office jurisdiction. In most of the cases, this is a country, with the exception of the European Patent Office (EPO) and the African Regional Intellectual Property Organization (ARIPO) where the countries of patent jurisdiction are under the control of the applicant (Hall 2017).

---

<sup>1</sup>Until 1970, what we currently know as WIPO was called *Bureaux Internationaux Réunis pour la Protection de la Propriété Intellectuelle* (French for *The United International Bureaux for the Protection of Intellectual Property*).



## 2.2 *Search and Retrieval on Patent Data*

The patent and monopolies offices, in their very early stages, were doing little more than registering, filing, and classifying the inventions.<sup>2</sup> The basic principles of patent examination were laid down by the adoption of the US Patent Act in 1836, principles which were soon adopted by other countries (1902 in France, 1877 in Germany) (Skolnik 1977). Already in the 1970s the information retrieval problem was an issue: with over 600,000 worldwide applications per year, a large number at that time, only partial retrieval solutions were available, most of them based on the classification systems (McDonnell 1969). While studying local clustering in full-text searches using local feedback, experiments were done on a small database of US patents (Attar and Fraenkel 1977). Attar and Fraenkel (1977) did an experiment that was a “technology survey”-like search on a set of 76 US patents. Two decades later a “prior art search” was performed on 13,747 US patents where the topics of the search were patents and citations were used to generate relevance assessments (Osborn et al. 1997).

In the last decades, research in IR methods for the IP domain has intensified. Workshops, conferences and evaluation tracks were organized in an effort to bring IR and IP communities together (see Iwayama et al. 2003; Kando and Leong 2000; Tait et al. 2010; Hanbury et al. 2010). The National Institute of Informatics (NII), Japan, initiated a series of workshops and evaluations using patent data as part of the NTCIR project (the NII Test Collections for IR Systems, currently renamed as the NII Testbeds and Community for Information access Research), focusing on Japanese and Chinese patents, and their translations into English.

In 2009, two further evaluation activities using patent data were launched: TREC-CHEM and CLEF-IP. TREC-CHEM ran from 2009 to 2011 and was organized as a chemical IR track in TREC (Text Retrieval Conference) addressing challenges in Chemical and Patent Information Retrieval (Lupu et al. 2009). The document collection used by TREC-CHEM was limited to chemical patent documents and chemical journal articles.

The purpose of the CLEF-IP track was to encourage and facilitate research in the area of multilingual patent retrieval by providing a large, clean data set for experimentation. The data set contains patents in three main European languages, patents published by the European Patent Office (EPO), as well as queries and associated relevance judgements.

---

<sup>2</sup>The first US classification system consisted of 16 classes in 1830.

### 2.3 *Characteristics of Patent Data*

We give a brief account of the main phases in the patenting process, establishing at the same time the basic patent notions and the different patent related aspects that are used throughout this chapter.

**Pre-application Phase** A person having developed an invention, will first write down a document describing the invention's background, a detailed description of it, and a set of claims that specify the extent of the protection sought. The level of detail of each of the document parts may vary depending on the patent office. The claims part of the application document is a legal text, therefore it is common to get the help of a patent attorney to draft it. This leads to the patent document having a mixture of writing styles, with the description of the invention being written in a narrative style, while the claims are written in a legal style (also called "attornish" or "patentese").

Before registering this document with a patent office, the inventor usually does what is called a "technology survey" of the existing technology in the area of his or her invention, the results of the search possibly triggering a change in the invention's specifications.

**Examination Phase** Upon registering the document with a patent office it becomes known as the "Patent Application Document" and receives an alphanumerical code that uniquely identifies it among other patent applications.

When a patent application is filed at a patent office, the application is given to patent professionals for examination. Each patent office follows different laws when deciding which claims to grant, but there is a set of worldwide common criteria that have to be fulfilled by any application before a patent can be granted (EPO 2018):

- **novelty:** the invention should not be previously known;
- **inventive step:** the invention should not be obvious for experts in the technological area of the invention;
- **realizable:** the invention can be manufactured by experts in the area.

The novelty check for an invention is done by performing a thorough search on the data collections available to the patent expert examining the patent application. The novelty search is the most time consuming and expensive part of the application examination. According to personal communications with various patent experts, the examination for novelty can take up to several weeks and even months, searches being repeated sometimes on different areas of the available databases, or with different sets and combinations of keywords. The result of a novelty search (also known as a "Prior Art Search") is a list of relevant documents stored into a "Search Report"; the relevant documents are called patent citations (note the different meaning of the word "citation" compared to academic publications). The citations

that an examiner found to be relevant to an application can be of three main types, which (in their order of relevance) are:

- citations that describe prior work but which do not destroy the novelty of the application (lower relevance);
- citations that, in combination with other citations, destroy the novelty of an application;
- citations which, taken alone, make a patent application not novel (high relevance).

**Granting and Opposition Phases** When the search report is created, a series of official communications between the applicant and the patent office take place. As an output of these communications claims are usually modified in order not to infringe existing patents. Quite often, patent applications are withdrawn.

When the patent office takes the decision to grant a patent, a “Granted Patent Document” is published. From this point on, for a certain amount of time (9 months at the EPO) oppositions to a granted patent may be filed to the patent office.

In this chapter we refer to the documents generated during the patenting phases as “patent documents.” One patent will, administratively, consist of several patent documents, like the Patent Application Document, the Search Report or the International Search Report, the Granted Patent Document.

### 2.3.1 Types of Patent Search

Depending on the type of information need and on the starting parameters of the search, the process used in finding relevant patents can differ from case to case and from one practitioner to another (Lupu and Hanbury 2013). A detailed description of the types of patent search can be found, for example, in Adams (2011), Alberts et al. (2011), Hunt et al. (2007) and should be differentiated from the IR task of searching in test collections that contain patents.

The search types that are typically performed in the three patent life-cycle phases above are:

- Pre-Application Search (technology survey) which is a search done by the inventor before filing for a patent application. The goal of the search is to identify existing knowledge (printed or not, including patents) which pertains to the invention.
- Novelty Search which aims to establish the novelty or the lack of novelty of an invention. This search can be performed both for filed patent applications or granted patents, as well as for inventions that were not yet filed.
- Patentability or Validity Search which is a search to identify prior art (that is previously published documents) that are relevant to the inventiveness of a patent application. Such searches may include novelty searches and are often carried out during the examination of a patent application.

### 2.3.2 Patent Data Is Administrative Data

During the patenting process a large number of documents are usually created, both by the patent office and by the applicant or her attorney. Communications to/from the patent office, application document amendments, registration of fee payments, and designating the states where the patent is valid are all examples of information that belong to the patent itself.

The general understanding of the patent concept is that, through its claims, it restricts other parties from exploiting the invention described in the respective granted patent. However, if we view patents as the complete set of documents generated during the patenting process, we immediately notice that patent data has a substantial administrative side. The administrative data includes, for example, application dates, addresses of the inventors and/or patent assignees, priority references, legal status, and so on. Of interest for the CLEF-IP tasks presented here are the patent classification system and the patent families.

**Patent Clustering by Families** In the current global economy, often enough after filing an initial patent application, inventors will pursue legal protection for their invention in additional countries of interest for them. Following the general patenting process, they will file subsequent applications at each patent office in the countries of interest referring to the original filing as the “priority claim”. Even though these applications may somewhat differ in content, depending on the patent laws in force at the various patent offices, it is obvious that, worldwide, patent content is often replicated. To assist patent practitioners with minimizing the necessary documents they might need to inspect, several methods to group ‘parallel’ patent documents were devised. The group of applications pertaining to the same invention is called a “patent family”.

There is no single definition of what a patent family is. Moreover, each provider of patent data constructs the patent families differently. For example, the EPO uses three types of patent family, while the WIPO additionally defines three further types (WIPO 2013). Nevertheless, as with the patent classification systems, the patent families are widely used when dealing with patent data.

**Patent Classification by Technological Areas** Patent classification systems are designed to categorize the patent documents by technological areas and sub-areas, using the technical features of the disclosed inventions. Several patent classification systems are in use, systems created both by patent offices and by private companies. The most well known are the International Patent Classification System (IPC),<sup>3</sup> the United States Patent Classification (USPC),<sup>4</sup> the F-term Japanese Classification System (Schneller 2002), or the Derwent Classification System.<sup>5</sup> Since January

---

<sup>3</sup>International Patent Classification (IPC) [www.wipo.int/classifications/ipc/en/](http://www.wipo.int/classifications/ipc/en/).

<sup>4</sup>United States Patent Classification [www.uspto.gov/patents/resources/classification/](http://www.uspto.gov/patents/resources/classification/).

<sup>5</sup>Derwent World Patents Index [clarivate.com/products/dwpi-reference-center/dwpi-classification-system/](http://clarivate.com/products/dwpi-reference-center/dwpi-classification-system/).

2013 the EPO and the USPTO (US Patents and Trademarks Office) use a joint classification system, the Cooperative Patent Classification system (CPC).<sup>6</sup>

In the early days of the patent system, patent classification systems were designed as a shelf-location tool for paper files (Adams 2000). Even today, these systems are manually maintained by experts and represent a ubiquitous resource for augmenting the query terms of on-line patent retrieval environments.

### 3 A Collection of European Patent Documents

One of our aims at the time we embarked on the CLEF-IP endeavour was to create a test collection fit for experimenting with patent data, a collection that faithfully mirrors the features and challenges of the data used in the actual working cycles of a patent professional. For this we use actual patent documents from the EPO and WIPO. These documents contain most of the information that is actively used by patent practitioners in their daily work with patent data.

The bulk of the collection's corpus is made of patent documents stored as XML files. In CLEF-IP, a patent consists of one or more XML files, one for each patent document that was available at the time of the collection creation. Since its first release in 2009, consecutive additions were made to the CLEF-IP test collection, so that it currently contains almost 3.5 million XML files corresponding to almost 1.5 million patents. These patents are an extract from the larger MAREC<sup>7</sup> collection which contains files representing over 19 million patents published at the EPO, USPTO, WIPO and JPO (Japan Patent Office) stored in a common normalized XML format. The main elements of the XML representations are shown in the simplified listing below:

```
<patent-document>
  <bibliographic-data> ... </bibliographic-data>
  <abstract> ... </abstract>
  <description> ... </description>
  <claims> ... </claims>
</patent-document>
```

The <abstract>, <description>, and <claims> elements store the textual content of the disclosed invention. These fields may occur more than once when, for example, both the English and the German versions of the abstract are stored in a patent document. Most of the patent text retrieval methods make use of the abstract, description and claims fields. The <bibliographic-data> element contains the administrative data related to a patent. In this XML element we will find the application and publication dates and references, family identifiers, the patent

---

<sup>6</sup>Cooperative Patent Classification (CPC) [www.cooperativepatentclassification.org/](http://www.cooperativepatentclassification.org/).

<sup>7</sup>The MAtrixware REsearch Collection. <http://ifs.tuwien.ac.at/imp/marec>.

classification symbols, inventors, assignees, postal addresses of the inventors and/or assignees, the invention title (in three languages), and the patent citations relevant to the invention in this document.

The CLEF-IP collection is limited to the MAREC patents published by the EPO, patents with application date earlier than 2002. The EPO patent documents published later were retained to form a *test and training topic pool* of approximately 500,000 patents, out of which we extracted training sets and topic sets for the CLEF-IP tasks (Graf and Azzopardi 2008).

In the corpus of European patent documents with application date prior to 2002, a high percentage of the patent documents refer to applications internationally filed under the Patent Cooperation Treaty (PCT 1970), also known as “EuroPCTs”. For these filings, the EPO does not republish the whole patent application, but only bibliographic entries that link to the original application published by the WIPO. Using text-based methods to retrieve such documents is problematic, and therefore, for these patent documents we added their WIPO equivalent to the CLEF-IP collection. Determining that the EuroPCT patent documents refer to a certain invention disclosed in a document published by WIPO is done by the patent family identifier which for the two documents must be the same. In this way, the collection became both larger and more realistic.

One of the most important features of the CLEF-IP corpus is its multilingualism. Patent applications to the EPO are written in one of the three official EPO languages (German, English, French), with the additional requirement that, once the decision to grant a patent is made, the claims section of the patent document must be submitted in all these three languages. Although the English language is over-represented in the CLEF-IP collection (see Table 1), not least due to the EuroPCT applications written in their large majority in English, the collection entails large amounts of content that is in German and French, making the collection suitable for carrying out multilingual retrieval experiments.

According to the specifics of each organized task, further chunks of data were added to the core CLEF-IP patent collection. One such data addition consisted of image files occurring in patents intended to support the concurrent use of textual and visual retrieval methods into one multimodal information retrieval method.

**Table 1** Document distributions in CLEF-IP

| 3.1 million documents |                     |             |
|-----------------------|---------------------|-------------|
| 14% WIPO documents    | 74% applications    | 67% English |
| 86% EPO documents     | 26% granted patents | 22% German  |
|                       |                     | 6% French   |
|                       |                     | 5% Unknown  |

## 4 The CLEF-IP Text Retrieval Tasks

There were five CLEF-IP evaluation cycles with a total of 7 tasks (Table 2). Some of the tasks were organised once only (e.g. the “Chemical Structure Recognition” task), others ran for 2 or 3 years in a row.

The “Prior Art Candidates” task (PAC, 2009–2011) required that, for a given patent application document (the “topic patent”), all patent documents relevant to the described invention are retrieved. The “Passage Retrieval (Starting from Claims)” task (PSG, 2012–2013) required that, given a patent application document and a selected subset of its claims, all patents that may invalidate these claims are retrieved, and, in addition, the concrete passages that do so are returned.

The “Patent Classification” task (CLS, 2010–2011) requested that a given patent document was classified according to the IPC classification symbols.

To solve these three tasks—PAC, PSG, and CLS—only text based analysis of the available CLEF-IP test collection files was necessary. Besides these text retrieval and classification tasks, and as part of the CLEF-IP campaign, further tasks that involved analysis of images in patents were organised between 2011 and 2013.

The “Image-based (Prior Art) Retrieval” task (IMG-PAC, 2011) asked the participants to retrieve relevant patents to the invention in a given topic patent, where, in addition to the text content in the XML patent documents, we provided the images that were attached to the patents. For more details on this task see (Piroi et al. 2011).

The “Image Classification” task (IMG-CLS, 2011) required that 1000 topic patent images (figures attached to patents) were classified into one of nine classes: drawing, chemical structure, program listing, gene sequence, flow chart, graph, mathematics, table, and symbol. No text analysis was necessary for this task.

The “Flowchart/Structure Recognition” task (2012–2013) and the “Chemical Structure Recognition” task (2012) didn’t necessitate text analysis either, as they required participants to extract content from patent images and store it into a predefined textual format in order to make it search-able by text-based IR methods.

Table 2 gives an overview of the CLEF-IP tasks and number of topics by the year they were organised. The last four tasks that involve image analysis are not the subject of this chapter, for more details we direct the reader to the references

**Table 2** CLEF-IP tasks, number of topics in the main topic sets, and year of their organisation

| Task/year   | 2009   | 2010 | 2011 | 2012 | 2013 |
|---|--------|------|------|------|------|
| Prior art candidates (PAC)                                | 10,000 | 2000 | 3973 |      |      |
| Passage retrieval (PSG)                                   |        |      |      | 105  | 149  |
| Patent classification (CLS)                               |        | 2000 | 3000 |      |      |
| Image-based retrieval (IMG-PAC) (Piroi et al. 2011)       |        |      | 211  |      |      |
| Image classification (IMG-CLS) (Piroi et al. 2011)        |        |      | 1000 |      |      |
| Flowchart/structure recognition (Piroi et al. 2012, 2013) |        |      |      | 100  | 747  |
| Chemical structure recognition (Piroi et al. 2012)        |        |      |      | 865  |      |

```
<topic>
  <num>EP-1222860-A2</num>
  <narr>Find all patents in the collection that
    potentially invalidate patent application
    EP-1222860-A2.</narr>
  <file>EP-1222860-A2.xml</file>
</topic>
```

**Fig. 1** Excerpt from the file with the list of topics in CLEF-IP PAC tasks

indicated in Table 2. In the following we detail the design of each CLEF-IP text-related task, the data used to extract topics and relevance judgments for the topics.

## 4.1 Topic Sources

The topics for each of the PAC, PSG, and CLS tasks consisted of an XML file corresponding to patent applications published between 2002 and 2008, selected from the *test and training topic pool*. In 2009 the topics were selected such that at least one highly relevant patent citation per topic was contained in the CLEF-IP collection. A further condition on topic selection, in 2009, was that, for a topic patent, the XML patent document is a Granted Patent Document which, according to the EPO regulations, provides the claims in the three EPO official languages (German, English, French).<sup>8</sup> With this decision we gave the task participants incentives to investigate cross-language retrieval methods already in the first CLEF-IP evaluation cycle.

In 2010 and 2011, to model the IP professional work procedures and rules more realistically, the topic patents are Patent Application Documents. We sampled the topic patents by their document language, by available citations within the CLEF-IP collection, and by their IPC class, such that each IPC class is equally represented in the final topic test set.<sup>9</sup> To further stimulate the research into cross-language patent retrieval methods, whenever possible, we selected topic patents where the language of the patent citation document was different from the language of the patent application document language (e.g. application document language is English, while the document language of a relevant patent citation in the search report is French or German).

The list of topics is stored as an XML file where the topic identifier is the patent number as assigned by the EPO (Fig. 1).

---

<sup>8</sup>The occurrence of multi-lingual content is a consequence of the Rule 71(3) of the European Patent Convention (EPO 1973) which states that granted patents must contain claims in the three official languages of the EPO.

<sup>9</sup>IPC classification represents the different domains of the patent applications: chemistry, textiles, mechanical engineering, physics, electricity, etc.



---

```

<tid>PSG-2</tid>
<tfile>EP-1445439-A1.xml</tfile>
<tfam-docs>FI-116479-B1.xml,FI-20030196-A.xml,FI-20030196-D0.xml</tfam-docs>
<tclaims>/patent-document/claims/claim[1] /patent-document/claims/claim[2]
/patent-document/claims/claim[3] /patent-document/claims/claim[4]</tclaims>

```

---

**Fig. 2** Excerpt from the file with the list of topics in CLEF–IP PSG tasks

Examining the EPO patent search reports closer, we immediately observe that, besides the list of patent citations relevant to a patent application, the reports detail which parts of a citation document (lines, columns, figures, etc.) are pertinent to which particular claims of the patent application. Therefore, in 2012 and 2013, we changed the PAC task formulation from ‘find relevant documents’ to ‘find relevant documents and mark in them the passages of interest to a given set of patent application claims’ (PSG). At the same time, although the basis for topic creation remained the same—actual patent application documents from the topic pool—the topics are now (sub)sets of claims in the patent application document, instead of the patent application document itself. It also allowed us to extract more than one topic (set of claims) out of one patent application document (Piroi et al. 2012, 2013). Figure 2 is an example of a topic in the CLEF–IP 2013 PSG list of topics file: Although the PSG topics contained only claims, it was allowed to use other parts of the topic’s application patent document for query generation. Moreover, in 2013, each topic contained also the reference to the patent document that constituted the priority claim document of the topic application document. Examiners at patent offices also have access to this kind of information related to new, incoming patent applications.

We note that, for each task and each year, the topic sets did not overlap. Similarly, for each of the three tasks and in each year, distinct sets of training topics were provided to the participants.

We conclude this subsection with a few remarks on the topic document’s language. In 2009, in addition to the main topic set where no restrictions on the document’s language were applied, three additional language specific tasks were created, where the topics in each of the three sets were documents in only one of the three EPO official language. In 2010, where no language specific tasks were organised, we did not impose restrictions on the document language when selecting the topics, which resulted in the obvious fact that the document language distribution in the topic set followed the document language distribution in the collection corpus (see Table 1). A consequence of this ‘natural’ language distribution was that methods using distinct algorithms for the different languages to process, index, and search the documents were not easy to qualitatively assess with respect to their language specific methods. We compensated for this in the following years where each third of the topic set contained documents written in one of the official EPO languages. The same is true for the training sets as well, where each EPO language was represented by a third of the topics.

## 4.2 *Relevance Assessments and Metrics*

Any organiser of an IR evaluation campaign faces the challenge of how to best obtain the ground truth for the topic test sets in order to be able to judge the quality of the submitted retrieval results. The big majority of the evaluation efforts (TREC, CLEF) use some form of document pooling from the submitted retrieval experiments, manually assessing the relevance of the documents in the pool by volunteer work (Spark-Jones and Van Rijsbergen 1975; Voorhees and Harman 2005). Recently, efficient pooling strategies have been proposed such that human effort may be reduced (Lipani et al. 2017). Still, obtaining humanly created relevance assessments is time-consuming and, in the case of patent evaluation, volunteers are difficult to find as costly expert knowledge is required (Roda et al. 2010). At the same time, because of strict regulations in logging their work, patent experts at patent offices do provide partial relevance assessments in the form of patent citations in the search reports. These relevance assessments are of high quality and, furthermore, at the EPO, the patent citations have relevance degrees assigned to them (see Sect. 2.3, Examination Phase).

However, using search reports as a source for relevance assessments gives an average of six relevant documents for a patent application document. This low number did not change over the years. In 1996/1997, in their experiment with patent retrieval, Osborn et al. found that their test collection also showed an average of six documents per query (Osborn et al. 1997). Nevertheless, we extracted relevance assessments from patent search reports following the general lines described in Graf and Azzopardi (2008). To increase the number of relevant documents we made use of patent families by creating an extended list of citations which includes the patent citations of the topic patent application document, the patent citations of the topic document's family members and the family members of the patent citation documents. After filtering out the patent citations that are not part of the CLEF-IP corpus, we reached an increase in the number of relevant documents by a factor of 7 (Roda et al. 2010).

As explained above, we used patent families to extract relevance assessments for the PAC topics. Obtaining the relevance assessments for the CLS task was straight forward: the IPC relevant classes were extracted from the classification assigned by the patent offices and present in the administrative part of the documents (the `<bibliographic-data>` XML field).

Extracting the relevance assessments for the PAC and the CLS tasks could be done automatically. The relevance files contain lists of *(topic, relevant document)* identifier pairs, where the identifiers referred to documents in the collection. The situation was more challenging for the PSG task, where we could not make use of patent families any more. In this task both the topics and the relevance assessments contain XPaths to the claims and relevant passages in the XML patent documents. The relevance files contain lists of *(topic, relevant document, relevant passage XPath)* identifier triples where the relevant document identifier refers to patent documents relevant to the topic, and the passage XPath identifies, within the relevant

| DOCUMENTS CONSIDERED TO BE RELEVANT |   |                   |
|-------------------------------------|---|-------------------|
| Category                            | Citation of document with indication, where appropriate, of relevant passages   | Relevant to claim |
| X                                   | WO 98 07379 A (LARSEN ERIC ;HOEGSETH SOLFRID (NO))<br>26 February 1998 (1998-02-26)                                       | 1-7,14,15         |
| Y                                   | * page 5, paragraph 1 - page 6, paragraph 2; figures 2,3 *<br>---   | 8-11              |
| X                                   | WO 01 26573 A (COHERENT INC)<br>19 April 2001 (2001-04-19)<br>* page 13, line 30 - page 15, line 16;<br>figure 3 *<br>--- | 1-3,7             |

**Fig. 3** Extract from a search report

document, the passage that is pertinent to the claims in the PSG topic. For the PSG task the relevant passage information was extracted manually by matching the passage indications in the search reports (Fig. 3) with the textual content of the patent documents in our corpus. When matched, we extracted the XPath of the identified content and saved them to a database. This process was time consuming, the main hurdle being comparing the PDF patent documents to which the search reports refer with the XML content of the document in the CLEF-IP collection. Therefore, the number of topics in the PSG test sets is low compared to the number of topics in the PAC and CLS tasks.

The measures reported for the PAC tasks are Precision and Recall at different cut-offs, MAP, nDCG (Järvelin and Kekäläinen 2002), and PRES (Magdy and Jones 2010a). For the CLS tasks we computed Precision, Recall and  $F_1$  at one classification code and at five classification codes (a patent may be classified into more than one IPC class). Since the PSG relevance assessments were triples, the evaluation for this task could be done on two levels: at the relevant document level and at the relevant passage (XPath) level. The evaluation at the document level measured a system's performance in retrieving whole relevant documents, very similar to the evaluations done in the PAC task, while the evaluation at the passage level targeted measuring the ranking quality of the passages in the relevant patent documents (Piroi et al. 2012). At the document level we maintained the computation of MAP, Recall and PRES measures. At the passage-level we assessed the systems' quality w.r.t. the relevance of the returned passages (XPath) by computing MAP and Precision scores for the retrieved passages grouped by relevant documents (MAP(D) and Precision(D)) and then averaging over the set of topics. These two document level measures carry similarities with the 'Relevant in Context' metrics of the INEX campaign (Kamps et al. 2008), but looking at sequences of XPath instead of sequences of characters (Piroi et al. 2012, Section 2.1).

## 5 Submissions and Results

For all CLEF-IP tasks, a *submission* (or *run*) consisted of a single text file with at most 1000 answers per topic. The most answers were given for the Prior Art Tasks, while the Patent Classification tasks required fewer answers per topic. With few variations, the format of the submissions followed the format used for the TREC submissions, which is a list of tuples containing at least the *topic identifier*, the retrieved *document* (and *passage* for the PSG tasks), the *rank* of the retrieved answer, and the *score* given by the retrieval system to the retrieved answer. Table 3 lists the groups that have submitted experiments to the PAC, PSG, and CLS tasks.

Generally, participants in the CLEF-IP evaluation benchmark have used off-the-shelf retrieval and classification engines (Indri/Lemur or Terrier engines, commonly available k-nearest neighbour algorithm implementations, support vector machines, SVM, or Winnow-like classifiers), choosing to tune these systems on the provided training sets. The better results, however, were obtained by those systems that put more effort into understanding and exploiting the patent specific data, like citations or classification symbols (Lopez and Romary 2009, 2010; Magdy and Jones 2010b; Mahdabi et al. 2011).

Some of the participants did experiments to determine which parts of the (topic) patent documents contribute most to improving retrieval results. These included selecting certain file parts to index, building separate indexes per document XML field, or boosting query terms extracted from certain parts of the topic files (Gobeill et al. 2009; Becks et al. 2010; Gobeill and Ruch 2012; Verberne and D'hondt 2011).

Given that each patent document could contain text in up to three languages, some participants chose to build separate indexes per language (Lopez and Romary 2009; Szarvas et al. 2009), while others generated one mixed-language index or used text fields only in one language discarding information given in the other languages (Correa et al. 2009; Toucedo and Losada 2009). Few participants made use of machine translations to obtain query terms in additional languages and applying them on the previously created collection indexes (Magdy and Jones 2010b). The granularity of the index varied, too, as some participants chose to concatenate all text fields into one index, while others indexed different fields separately. In addition, several specific indexes like phrase or passage indexes, concept indexes and IPC indexes were used (Magdy et al. 2009; Wanagiri and Adriani 2010; Szarvas et al. 2009). A more detailed analysis of the indexing methods and of the retrieval approaches used in the 2009 and 2010 evaluation labs can be found in Piroi and Zenz (2011).

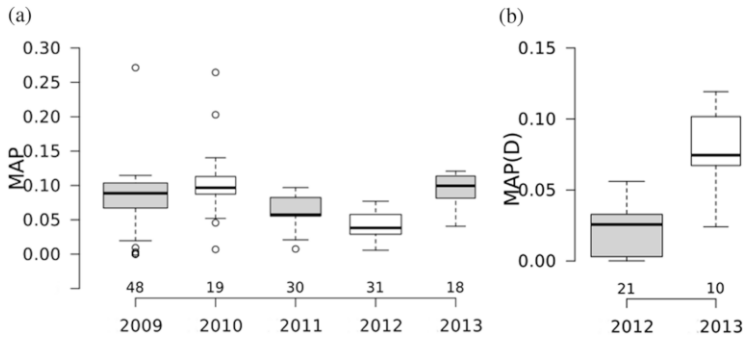
As the task topics were complete documents, with several pages of texts, extracting appropriate queries from the topic document has been investigated by several participating teams (Graf et al. 2009; Becks et al. 2009).

The IPC classification codes were the part of the <bibliographic-data> that was exploited the most and was used either as a post-processing filter, as part of the query, or to pre-select smaller sets of patents to search in Gobeill et al. (2009), Szarvas et al. (2009), Eiselt and Oberreuter (2013), Lopez and Romary (2009),

**Table 3** Teams that participated in the text-based retrieval and classification CLEF-IP tasks

| Team  | 2009           | 2010                       | 2011                       | 2012           | 2013           |
|---|----------------|----------------------------|----------------------------|----------------|----------------|
| BiTeM, Service of Medical Informatics, Geneva Univ. Hospitals CH                | PAC            | PAC<br>CLS                 |                            | PSG            |                |
| Centrum Wiskunde & Informatica - Interactive Information Access NL              | PAC            | PAC<br>CLS                 |                            |                |                |
| Chemnitz Univ. of Technology, Dept. of Computer Science DE                      |                |                            | PAC                        | PSG            |                |
| Dublin City Univ., School of Computing UK                                       | PAC            | PAC                        |                            |                |                |
| Geneva Univ., Centre Universitaire d'Informatique, SimpleShift CH               | PAC            | CLS                        |                            | PSG            |                |
| Gerogetown Univ., Dept. of Computer Science US                                  |                |                            |                            |                | PSG            |
| Glasgow Univ. - IR Group Keith UK   | PAC            |                            |                            |                |                |
| Hewlett-Packard Labs, Russia RU   |                |                            | PAC                        |                |                |
| Humboldt Univ., Dept. of German Language and Linguistics DE                     | PAC            | PAC<br>CLS                 |                            |                |                |
| Industrial Property Documentation Dept., JSI Jouve FR                           |                | CLS                        |                            |                |                |
| Innovandio S.A. CL  |                |                            |                            |                | PSG            |
| Inria FR  | PAC            | PAC<br>CLS                 |                            |                |                |
| SIEL, International Institute of Information Technology IN                      |                |                            | PSG                        |                |                |
| LCI – Institut National des Sciences Appliqu'ees de Lyon FR                     |                | CLS                        |                            |                |                |
| Radboud Univ. Nijmegen NL   | PAC            | CLS                        | CLS                        |                |                |
| Santiago de Compostela Univ., Dept. Electronica y Computacion ES                | PAC            |                            |                            |                |                |
| Spinque B.V. NL   |                | PAC<br>CLS                 | PAC                        |                |                |
| Swedish Institute of Computer Science SE  | PAC            |                            |                            |                |                |
| Technical Univ. Darmstadt, Dept. of CS, Ubiquitous Knowledge Processing Lab DE  | PAC            |                            |                            |                |                |
| Technical Univ. Valencia, Natural Language Engineering ES                       | PAC            |                            |                            |                |                |
| UNED - E.T.S.I. Informatica, Dpto. Lenguajes y Sistemas Informaticos, Madrid ES |                | PAC                        |                            |                |                |
| Univ. Indonesia, Information Retrieval Group ID                                 |                | PAC                        |                            |                |                |
| Univ. "Alexandru Ioan Cuza", Iași RO  | PAC            | PAC                        |                            |                |                |
| Univ. of Hildesheim, Information Science DE                                     | PAC            | PAC                        | PAC                        | PSG            |                |
| Univ. of Lugano CH  |                |                            | PAC                        | PSG            |                |
| Univ. of Macedonia, Dept. of Applied Informatics, Thessaloniki GR               |                |                            |                            | PSG            | PSG            |
| Univ. of Neuchatel, Computer Science CH   | PAC            |                            |                            |                |                |
| Univ. of Tampere - Info Studies & Interactive Media FI                          | PAC            |                            |                            |                |                |
| Univ. of Wolverhampton, School of Technology UK                                 |                |                            |                            | PSG            |                |
| Vienna Univ. of Technology, IFS AT  |                |                            | PAC                        | PSG            | PSG            |
| WISEnut Ltd. KR   |                |                            | PAC<br>CLS                 |                |                |
| <b>Total runs:</b>  | <b>PAC: 48</b> | <b>PAC: 25<br/>CLS: 27</b> | <b>PAC: 30<br/>CLS: 25</b> | <b>PSG: 31</b> | <b>PSG: 18</b> |

The gray shading are a means to distinguish the consecutive table lines and has no other meaning



**Fig. 4** Summary of MAP scores in the PAC and PSG CLEF-IP tasks. (a) MAP scores for the PAC tasks. (b) MAP(D) scores for the PSG tasks

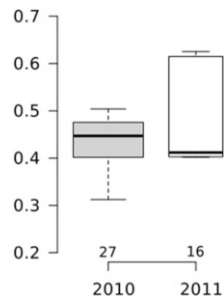
Giachanou et al. (2013). The patent citation information stored in the document set of the collection was exploited less in 2009, with more groups using this metadata in the following years. Other very patent-specific information, like filing dates, applicant and inventor names and/or countries, was rarely used.

To give an idea of the score ranges achieved by retrieval systems participating in the Prior Art tasks, we show in Fig. 4 box plot summaries of the submitted run scores for mean average precision, MAP, and passage mean average precision, MAP(D), for each year where these tasks ran.<sup>10</sup> The numbers just above the years on the  $x$ -axis show the number of valid runs submitted and evaluated in the respective year. The main take away message from observing the box plots in Fig. 4 is that most IR strategies, however different in their design and methods, are equally inefficient in tackling the patent retrieval tasks. The positive outliers in these figures are, in fact, scores obtained by IR systems that integrated patent domain expertise in their design. An examples of such expertise is the query expansion with terms that do not necessarily occur in the topic patent document, but are extracted from the test collection by analysing IPC related information and/or the content of patent citations. It is also clear, from this figure, that Passage Retrieval in the patent domain, as defined by the PSG task, is an even more difficult retrieval problem.

The classification of patent documents proved to be an easier challenge than finding prior art using IR methods. This is reflected in the scores obtained by the participants' submissions. These are shown in Fig. 5 which summarises the  $F_1$  values obtained by the experiments submitted to the CLS tasks in 2010 and 2011.

Submissions to the Classification task either used text classifiers only, like kNN or Winnow type neural networks (Derieux et al. 2010; Guyot et al. 2010; D'hondt et al. 2011), or chose a solution implementing systems similar to text retrieval

<sup>10</sup>Note that the scores between years cannot be directly compared, as each lab year came with a new set of test topics.

**Fig. 5** CLS tasks  $F_1$  scores

that returned the IPC codes as results, or combined classification and text retrieval (Teodoro et al. 2010; Derieux et al. 2010).

All data related to the CLEF-IP evaluation campaign (collection, topics, scripts, documentation, etc.) can be downloaded from the CLEF-IP website.<sup>11</sup> Detailed descriptions of the systems that participated in the CLEF-IP tasks can be found in the CLEF workshop notes available on the CLEF Initiative website<sup>12</sup> and on the CLEF-IP website.

## 6 Closing Remarks

We have presented in this chapter the development of the CLEF-IP benchmarking activity for patent text retrieval over a period of 5 years. It advanced from a simply formulated retrieval task to organizing more elaborated tasks that cover specific pieces of the Intellectual Property practitioners' daily work-flow.

At the end of the CLEF-IP evaluation campaign, it is clear to us that successful information retrieval in the patent domain involves at least well thought-out adjustments to the currently used retrieval and text mining systems to take into account the specificities of the patent domain. In general, retrieval results do not come close to the expectations of patent experts. One reason for this is that transferring the know-how of IP professionals to the IR research community is a complex undertaking. An example of such patent domain expertise which was insufficiently treated by IR researchers is language obfuscation. A method used rather often by patent applicants, language obfuscation employs vague and over-broad terms for otherwise very concrete concepts.

Even though the CLEF-IP campaign is no longer running, there is a huge potential to use the data and realistic patent search tasks resulting from the CLEF-IP campaign to develop innovative solutions in the patent information

<sup>11</sup>CLEF-IP: Retrieval in the Intellectual Property Domain. <http://ifs.tuwien.ac.at/~clef-ip/>.

<sup>12</sup>The CLEF Initiative (Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum). <http://www.clef-initiative.eu/>.

retrieval domain. The CLEF-IP tasks described in this chapter are focused on text-oriented information retrieval. There remains however extensive work to be done on improving the use of non-textual patent data in patent search. Early steps in this direction were done by the organisation of additional tasks, where the CLEF-IP test collection was augmented with data sets pertinent to non-textual patent content: flowcharts, chemical structures, images.

Another important aspect of patent retrieval, which was not addressed by the CLEF-IP campaign, is that information search is session based: the final list of relevant documents is the result of several search queries, possibly building on each other. Both these research directions need sustained support from the IP community.

Undertakings like TREC-CHEM, CLEF-IP, NTCIR workshop series are ambitious from at least two points of view. On one side, by interfacing with patent practitioners, these evaluation activities can be used to showcase advances in IR methods, methods that should easily be adaptable to the IP domain, and facilitate their daily need for specific information needs, allowing them to explore the patent data in novel ways. On the other side, such evaluation campaigns repeatedly bring to the attention of academic IR researchers the fact that there exists a large body of technological know-how, namely patent databases. The CLEF-IP benchmark contributed to creating a picture of the search result quality the IR methods deliver when faced with an information need like the one represented by the patent novelty search (i.e. finding relevant patents for a given patent application). The availability of patent-based test collections has triggered research in various IR areas, an inventory of the latest IP-relevant studies being also presented in Lupu and Hanbury (2013) and Lupu et al. (2017).

**Acknowledgements** We give our thanks to the following people:

- the advisory board members which helped shape the evaluation lab in its early years: Gianni Amati, Atsushi Fujii, Makoto Iwayama, Kalervo Järvelin, Noriko Kando, Javier Pose Rodríguez, Mark Sanderson, Henk Thomas, Anthony Trippe, Christa Womser-Hacker
- the numerous patent experts that helped us understand many of the patent system's subtleties
- previous CLEF-IP organizers and co-organizers: Giovanna Roda, John Tait, Veronika Zenz, Mihai Lupu, Igor Filippov, Walid Magdy, Alan P. Sexton
- the participating teams who submitted such a variety of solutions to the proposed tasks (see CLEF-IP workshop notes).

CLEF-IP was supported along the years by Matrixware GmbH, Vienna and Information Retrieval Facility, Vienna as first data and infrastructure provider, by the PROMISE, EU Network of Excellence (FP7-258191), and FFG FIT-IT Impex project (No. 825846).

## References

- Adams S (2000) Using the International Patent Classification in an online environment. *World Patent Information* 22(4):291-300
- Adams SR (2011) *Information sources in patents*, 3rd edn. K.G. Saur, Munich



- Alberts D, Yang CB, Fobare-DePonio D, Koubek K, Robins S, Rodgers M, Simmons E, DeMarco D (2011) Introduction to patent searching. In: Lupu M, Mayer K, Tait J, Trippe AJ (eds) *Current challenges in patent information retrieval, the information retrieval series*, vol 29. Springer, Berlin, pp 3–43. [http://dx.doi.org/10.1007/978-3-642-19231-9\\_1](http://dx.doi.org/10.1007/978-3-642-19231-9_1)
- Anthon C (1841) *A classical dictionary: containing an account of the principal proper names mentioned in ancient authors, and intended to elucidate all the important points connected with the geography, history, biography, mythology, and fine arts of the Greeks and Romans together with an account of coins, weights, and measures, with tabular values of the same*. Harper & Bros
- Attar R, Fraenkel AS (1977) Local feedback in full-text retrieval systems. *J ACM* 24(3):397–417. <http://doi.acm.org/10.1145/322017.322021>
- Becks D, Womser-Hacker C, Mandl T, Kölle R (2009) Patent retrieval experiments in the context of the CLEF IP Track 2009. In: Borri F, Nardi A, Peters C, Ferro N (eds) (2009) *CLEF 2009 working notes*. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1175/>
- Becks D, Mandl T, Womser-Hacker C (2010) Phrases or terms? The impact of different query types. In: Braschler M, Harman DK, Pianta E, Ferro N (eds) (2010) *CLEF 2010 working notes*. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1176/>
- Correa S, Buscaldi D, Rosso P (2009) NLEL-MAAT at CLEF-IP. In: Borri F, Nardi A, Peters C, Ferro N (eds) (2009) *CLEF 2009 working notes*. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1175/>
- Derieux F, Bobeica M, Pois D, Raysz JP (2010) Combining semantics and statistics for patent classification. In: Braschler M, Harman DK, Pianta E, Ferro N (eds) (2010) *CLEF 2010 working notes*. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1176/>
- D'hondt E, Verberne S, Alink W, Cornacchia R (2011) Combining document representations for prior-art retrieval. In: Petras V, Forner P, Clough P, Ferro N (eds) (2011) *CLEF 2011 Working Notes*, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1177/>
- Eiselt A, Oberreuter G (2013) The simpler the better - Retrieval Model comparison for Prior-Art Search in Patents at CLEF-IP 2013. In: Forner P, Navigli R, Tufis D, Ferro N (eds) (2013) *CLEF 2013 working notes*. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- EPO (1973) *European Patent Convention (EPC), Implementing Regulations, Examination Procedure*. <http://www.epo.org/law-practice/legal-texts/html/epc/2016/e/r71.html>. Accessed Dec 2018
- EPO (2018) *Guidelines for Examination in the European Patent Office*. Directorate Patent Law 5.2.1. <http://www.epo.org/law-practice/legal-texts/guidelines.html>. Accessed Dec 2018
- Frietsch R, Schmoch U, Looy B, Walsh P, Devroede R, Du Plessis M, Jung T, Meng Y, Neuhäusler P, Peeters B, Schubert T (2010) *The value and indicator function of patents*. Expertenkommission Forschung und Innovation (EFI) Studien zum deutschen Innovationssystem 15(15-2010)
- Galasso A, Schankerman M (2013) *Do patents help or hinder innovation?* World Economic Forum. <https://www.weforum.org/agenda/2013/05/do-patents-help-or-hinder-innovation>. Accessed Dec 2018
- Giachanou A, Salampasis M, Satratzemi M, Samaras N (2013) Report on the CLEF-IP 2013 experiments: multilayer collection selection on topically organized patents. In: Forner P, Navigli R, Tufis D, Ferro N (eds) (2013) *CLEF 2013 working notes*. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Gobeill J, Ruch P (2012) *BiTeM site report for the Claims to Passage task in CLEF-IP 2012*. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) (2012) *CLEF 2012 working notes*. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>

- Gobeill J, Theodoro D, Ruch P (2009) Exploring a wide range of simple pre and post processing strategies for patent searching in CLEF IP 2009. In: Borri F, Nardi A, Peters C, Ferro N (eds) (2009) CLEF 2009 working notes. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1175/>
- Graf E, Azzopardi L (2008) A methodology for building a patent test collection for prior art search. In: Proceedings of the second international workshop on evaluating information access (EVIA)
- Graf E, Azzopardi L, van Rijsbergen K (2009) Automatically generating queries for prior art search. In: Borri F, Nardi A, Peters C, Ferro N (eds) (2009) CLEF 2009 working notes. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1175/>
- Guyot J, Benzineb K, Falquet G (2010) myClass: a mature tool for patent classification. In: Braschler M, Harman DK, Pianta E, Ferro N (eds) (2010) CLEF 2010 working notes. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1176/>
- Hall BH (2017) Patents. Palgrave, Macmillan, pp 1–9. [https://doi.org/10.1057/978-1-349-95121-5\\_1393-2](https://doi.org/10.1057/978-1-349-95121-5_1393-2)
- Hanbury A, Zenz V, Berger H (2010) 1st international workshop on advances in patent information retrieval (AsPIRe'10). SIGIR Forum 44(1):19–22. <http://doi.acm.org/10.1145/1842890.1842893>
- Hunt D, Nguyen L, Rodgers M (2007) Patent searching: tools & techniques. Wiley, New York
- Iwayama M, Fujii A, Kando N, Marukawa Y (2003) An empirical study on retrieval models for different document genres: patents and newspaper articles. In: Proceedings of the 26th international ACM SIGIR conference on research and development in information retrieval. ACM, New York, SIGIR '03, pp 251–258
- Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst 20(4):422–446. <http://doi.acm.org/10.1145/582415.582418>
- Kamps J, Peheveski J, Kazai G, Lalmas M, Robertson S (2008) INEX 2007 evaluation measures. In: Focused access to XML documents, 6th international workshop of the initiative for the evaluation of XML retrieval, INEX 2007, Dagstuhl Castle, Germany, December 17–19, 2007. Selected papers. Lecture notes in computer science, vol 4862. Springer, Berlin, pp 24–33
- Kando N, Leong MK (2000) Workshop on patent retrieval (SIGIR 2000 Workshop Report). SIGIR Forum 34(1):28–30
- Kumagai KI (2005) History of Japanese Industrial Property System. [http://www.jpo.go.jp/torikumi\\_e/kokusai\\_e/training/textbook/pdf/History\\_of\\_Japanese\\_Industrial\\_Property\\_System\(2005\).pdf](http://www.jpo.go.jp/torikumi_e/kokusai_e/training/textbook/pdf/History_of_Japanese_Industrial_Property_System(2005).pdf). Accessed Feb 2018
- Lipani A, Palotti J, Lupu M, Piroi F, Zuccon G, Hanbury A (2017) Fixed-cost pooling strategies based on IR evaluation measures. In: Advances in information retrieval - 39th European conference on IR research, ECIR 2017, Aberdeen, April 8–13, 2017, Proceedings, pp 357–368. [https://doi.org/10.1007/978-3-319-56608-5\\_28](https://doi.org/10.1007/978-3-319-56608-5_28)
- Lopez P, Romary L (2009) Multiple retrieval models and regression models for prior art search. In: Borri F, Nardi A, Peters C, Ferro N (eds) (2009) CLEF 2009 working notes. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1175/>
- Lopez P, Romary L (2010) Experiments with citation mining and key-term extraction for prior art search. In: Braschler M, Harman DK, Pianta E, Ferro N (eds) (2010) CLEF 2010 working notes. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1176/>
- Lupu M, Hanbury A (2013) Patent retrieval. Found Trends IR 7:1–97
- Lupu M, Huang J, Zhu J, Tait J (2009) TREC-CHEM: large scale chemical information retrieval evaluation at TREC. SIGIR Forum 43(2):63–70
- Lupu M, Mayer K, Kando N, Trippe A (2017) Current challenges in patent information retrieval, 2nd edn. Springer, Berlin. <https://doi.org/10.1007/978-3-662-53817-3>
- Magdy W, Jones G (2010a) PRES: a score metric for evaluating recall-oriented information retrieval applications. In: Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval, ACM, New York, SIGIR '10, pp 611–618. <http://doi.acm.org/10.1145/1835449.1835551>

- Magdy W, Jones GJF (2010b) Examining the robustness of evaluation metrics for patent retrieval with incomplete relevance judgements. In: Braschler M, Harman DK, Pianta E, Ferro N (eds) (2010) CLEF 2010 working notes. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1176/>
- Magdy W, Leveling J, Jones G (2009) DCU at CLEF-IP 2009: exploring standard IR techniques on patent retrieval. In: Borri F, Nardi A, Peters C, Ferro N (eds) (2009) CLEF 2009 working notes. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1175/>
- Mahdabi P, Andersson L, Hanbury A, Crestani F (2011) Report on the CLEF-IP 2011 experiments: exploring patent summarization. In: Petras V, Forner P, Clough P, Ferro N (eds) (2011) CLEF 2011 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1177/CLEF2011wn-CLEF-IP-MahdabiEt2011.pdf>
- May C (2010) The Venetian moment: new technologies, legal innovation and the institutional origins of intellectual property. *Prometheus* 20:159–179. <http://www.tandfonline.com/doi/full/10.1080/08109020210138979>
- McDonnell P (1969) Technical information management in the U.S. patent office. *J Chem Doc* 9(5):220–224
- Mossoff A (2007) Who cares What Thomas Jefferson thought about patents - reevaluating the patent privilege in historical context. *Cornell Law Rev* 92(5):953. <https://scholarship.law.cornell.edu/clr/vol92/iss5/2>
- Osborn M, Strzalkowski T, Marinescu M (1997) Evaluating document retrieval in patent database: a preliminary report. In: Proceedings of the 6th international conference on information and knowledge management. ACM, New York, pp 216–221. <http://doi.acm.org/10.1145/266714.266899>
- PCT (1970) Patent Cooperation Treaty. <http://www.wipo.int/pct/en/treaty/about.html>. Accessed Aug 2015
- Pfaller W (2013a) Bergrecht, Monopole, Privilegien. <http://www.wolfgang-pfaller.de/berg.htm>. Accessed Feb 2018
- Pfaller W (2013b) Schon die alten Griechen .... <http://www.wolfgang-pfaller.de/sybaris.htm>. Accessed Feb 2018
- Piroi F, Zenz V (2011) Evaluating information retrieval in the intellectual property domain: the CLEF-IP campaign. In: Lupu M, Mayer K, Tait J, Trippe AJ (eds) Current challenges in patent information retrieval, the information retrieval series, vol 29. Springer, Berlin, pp 87–108. [http://dx.doi.org/10.1007/978-3-642-19231-9\\_4](http://dx.doi.org/10.1007/978-3-642-19231-9_4)
- Piroi F, Lupu M, Hanbury A, Zenz V (2011) CLEF-IP 2011: retrieval in the intellectual property domain. In: Petras V, Forner P, Clough P, Ferro N (eds) (2011) CLEF 2011 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1177/>
- Piroi F, Lupu M, Hanbury A, Sexton AP, Magdy W, Filippov IV (2012) CLEF-IP 2012: retrieval experiments in the intellectual property domain. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) (2012) CLEF 2012 working notes. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Piroi F, Lupu M, Hanbury A (2013) Overview of CLEF-IP 2013 Lab - information retrieval in the patent domain. In: Forner P, Müller H, Paredes R, Rosso P, Stein B (eds) Information access evaluation meets multilinguality, multimodality, and visualization. Proceedings of the fourth international conference of the CLEF initiative (CLEF 2013). Lecture Notes in Computer Science (LNCS), vol 8138. Springer, Heidelberg
- Rich GS (1993) Are Letters Patent Grants of Monopoly? *Western New England Law Review* 15. <https://core.ac.uk/display/76563380>
- Roda G, Tait J, Piroi F, Zenz V (2010) CLEF-IP 2009: retrieval experiments in the intellectual property domain. In: Peters C, Nunzio GD, Kurimo M, Mostefa D, Penas A, Roda G (eds) Multilingual information access evaluation I. Text retrieval experiments 10th workshop of the cross-language evaluation forum, CLEF 2009, vol 6241. Springer, Berlin, pp 385–409
- Schneller I (2002) Japanese File Index Classification and F-terms. *World Patent Inf* 24(3):197–201

- Skolnik H (1977) Historical aspects of patent systems. *J Chem Inf Comput Sci* 17(3):119–121. <https://pubs.acs.org/doi/abs/10.1021/ci60011a002>
- Spark-Jones K, Van Rijsbergen C (1975) Report on the need for and provision of an ‘ideal’ information retrieval test collection. 5266, Computer Laboratory, Univ. Cambridge. [http://sigir.org/files/museum/pub-14/pub\\_14.pdf](http://sigir.org/files/museum/pub-14/pub_14.pdf)
- Szarvas G, Herbert B, Ggurevych I (2009) Prior art search using international patent classification codes and all-claims-queries. In: Borri F, Nardi A, Peters C, Ferro N (eds) (2009) CLEF 2009 working notes. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1175/>
- Tait J, Harris C, Lupu M (eds) (2010) PaIR ‘10: proceedings of the 3rd international workshop on patent information retrieval. ACM, New York
- Teodoro D, Gobeill J, Pasche E, Vishnyakova D, Ruch P, Lovis C (2010) Automatic prior art searching and patent encoding at CLEF-IP’10. In: Braschler M, Harman DK, Pianta E, Ferro N (eds) (2010) CLEF 2010 working notes. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1176/>
- Toucedo C, Losada D (2009) University of Santiago de Compostella at CLEF-IP’09. In: Borri F, Nardi A, Peters C, Ferro N (eds) (2009) CLEF 2009 working notes. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1175/>
- Verberne S, D’hondt E (2011) Patent classification experiments with the Linguistic Classification System LCS in CLEF-IP 2011. In: Petras V, Forner P, Clough P, Ferro N (eds) (2011) CLEF 2011 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1177/>
- Voorhees EM, Harman DK (2005) TREC: experiment and evaluation in information retrieval (digital libraries and electronic publishing). The MIT Press, Cambridge
- Wanagiri M, Adriani M (2010) Prior art retrieval using various patent document fields contents. In: Braschler M, Harman DK, Pianta E, Ferro N (eds) (2010) CLEF 2010 working notes. CEUR workshop proceedings (CEUR-WS.org), ISSN 1613-0073. <http://ceur-ws.org/Vol-1176/>
- WIPO (2013) Handbook on industrial property information and documentation. Part 8: terms and abbreviations concerning industrial property information and documentation. [http://www.wipo.int/standards/en/part\\_08.html](http://www.wipo.int/standards/en/part_08.html). Accessed Jan 2019

# Biodiversity Information Retrieval Through Large Scale Content-Based Identification: A Long-Term Evaluation



**Alexis Joly, Hervé Goëau, Hervé Glotin, Concetto Spampinato, Pierre Bonnet, Willem-Pier Vellinga, Jean-Christophe Lombardo, Robert Planqué, Simone Palazzo, and Henning Müller**

**Abstract** Identifying and naming living plants or animals is usually impossible for the general public and often a difficult task for professionals and naturalists. Bridging this gap is a key challenge towards enabling effective biodiversity information retrieval systems. This taxonomic gap was actually already identified as one of the main ecological challenges to be solved during the Rio de Janeiro United Nations “Earth Summit” in 1992. Since 2011, the LifeCLEF challenges conducted in the context of the CLEF evaluation forum have been boosting and evaluating the advances in this domain. Data collections with an unprecedented volume and diversity have been shared with the scientific community to allow repeatable and

---

A. Joly (✉)

Inria, LIRMM, Montpellier, France  
e-mail: [alexis.joly@inria.fr](mailto:alexis.joly@inria.fr)

H. Goëau · P. Bonnet

CIRAD, UMR AMAP, Montpellier Cedex 5, France  
e-mail: [herve.goeau@cirad.fr](mailto:herve.goeau@cirad.fr); [pierre.bonnet@cirad.fr](mailto:pierre.bonnet@cirad.fr)

H. Glotin

Université de Toulon, Aix Marseille Univ, CNRS, LIS, DYNI Team, Marseille, France  
e-mail: [herve.glotin@univ-tln.fr](mailto:herve.glotin@univ-tln.fr)

C. Spampinato · S. Palazzo

University of Catania, Catania, Italy  
e-mail: [cspampin@dieci.unict.it](mailto:cspampin@dieci.unict.it); [simone.palazzo@dieci.unict.it](mailto:simone.palazzo@dieci.unict.it)

W.-P. Vellinga · R. Planqué

Xeno-Canto Foundation, Amsterdam, The Netherlands  
e-mail: [wp@xeno-canto.org](mailto:wp@xeno-canto.org); [r.planque@vu.nl](mailto:r.planque@vu.nl)

J.-C. Lombardo

Inria, LIRMM, Montpellier, France  
e-mail: [jean-christophe.lombardo@inria.fr](mailto:jean-christophe.lombardo@inria.fr)

H. Müller

HES-SO, Sierre, Switzerland  
e-mail: [henning.mueller@hevs.ch](mailto:henning.mueller@hevs.ch)

long-term experiments. This paper describes the methodology of the conducted evaluation campaigns as well as providing a synthesis of the main results and lessons learned along the years.

## 1 Introduction

Identifying organisms is a key for accessing information related to the uses and ecology of species. This is an essential step in recording any specimen on earth to be used in ecological studies. Unfortunately, this is difficult to achieve due to the level of expertise necessary to correctly record and identify living organisms (for instance plants are one of the most difficult groups to identify with an estimated number of 400,000 species). This *taxonomic gap* has been recognized since the Rio Conference of 1992, as one of the major obstacles to the global implementation of the Convention on Biological Diversity. Among the diversity of methods used for species identification, Gaston and O'Neill (2004) discussed in 2004 the potential of automated approaches typically based on machine learning and multimedia data analysis. They suggested that, if the scientific community is able to (1) overcome the production of large training datasets, (2) more precisely identify and evaluate the error rates, (3) scale up automated approaches, and (4) detect novel species, it will then be possible to initiate the development of a generic automated species identification system that could open up vistas of new opportunities for theoretical and applied work in biological and related fields. Since the question raised by Gaston and O'Neill (2004), *automated species identification: why not?*, a lot of work has been done on the topic (e.g. Lee et al. 2004; Cai et al. 2007; Trifa et al. 2008; Towsey et al. 2012; Glotin et al. 2013a,b; Joly et al. 2014b) and it is still attracting much research today, in particular on deep learning techniques. In parallel to the emergence of automated identification tools, large social networks dedicated to the production, sharing and identification of multimedia biodiversity records have increased in recent years. Some of the most active ones like eBird<sup>1</sup> (Sullivan et al. 2014), iNaturalist,<sup>2</sup> iSpot (Silvertown et al. 2015), Xeno-Canto<sup>3</sup> or Tela Botanica.<sup>4</sup> SABIOD and EADM CNRS<sup>5</sup> federations on machine learning for bioacoustics (respectively initiated in the US for the two first ones, and in Europe for the others), federate hundreds of thousands of active members, producing millions of observations each year. Noticeably, PI@ntNet was the first initiative attempting to combine the force of social networks with automated identification tools (Joly et al. 2014b) through the release of a mobile application and collaborative

---

<sup>1</sup><http://ebird.org/content/ebird/>.




<sup>2</sup><http://www.inaturalist.org/>.

<sup>3</sup><http://www.xeno-canto.org/>.

<sup>4</sup><http://www.tela-botanica.org/>.

<sup>5</sup><http://sabiiod.org>.

validation tools. As a proof of their increasing reliability, most of these networks have started to contribute to global initiatives on biodiversity, such as the Global Biodiversity Information Facility (GBIF<sup>6</sup>) which is the largest and most recognized one. Nevertheless, this explicitly shared and validated data is only the tip of the iceberg. The real potential lies in the automatic analysis of the millions of raw observations collected every year through a growing number of devices but for which there is no human validation at all. The performance of state-of-the-art multimedia analysis and machine learning techniques on such raw data (e.g., mobile search logs, soundscape audio recordings, wild life webcams, etc.) is still not well understood and is far from reaching the requirements of an accurate generic biodiversity monitoring system. Most existing research before LifeCLEF actually considered only a few dozen or up to hundreds of species, often acquired in well-controlled environments (Goëau et al. 2011a; Nilsback and Zisserman 2008; Kumar et al. 2012). On the other hand, the total number of living species on earth is estimated to be around 10 K for birds, 30 K for fish, 400 K for flowering plants (cf. State of the World's Plants 2017<sup>7</sup>) and more than 1.2 M for invertebrates (Baillie et al. 2004). To bridge this gap, it is required to boost research on large-scale datasets and real-world scenarios. In order to evaluate the performance of automated identification technologies in a sustainable and repeatable way, the LifeCLEF<sup>8</sup> research platform was created in 2014 as a continuation of the plant identification task (Goëau et al. 2013b) that was run within the ImageCLEF lab<sup>9</sup> the 3 years before (Goëau et al. 2011a, 2012a, 2013a). LifeCLEF enlarged the evaluated challenge by considering birds and marine animals in addition to plants, and audio and video contents in addition to images. More concretely, the lab is organized around three tasks:

-  **PlantCLEF**: an image-based plant identification task making use of Pl@ntNet collaborative data, Encyclopedia of Life' data, and Web data
-  **BirdCLEF**: an audio recordings-based bird identification task making use of Xeno-canto collaborative data
-  **SeaCLEF**: a video and image-based identification task dedicated to sea organisms (making use of submarine videos and aerial pictures).

As described in more detail in the following sections, each task is based on big and real-world data and the measured challenges are defined in collaboration with biologists and environmental stakeholders so as to reflect realistic usage scenarios.

---

<sup>6</sup><http://www.gbif.org/>.

<sup>7</sup><https://stateoftheworldsplants.com/>.

<sup>8</sup><http://www.lifeclef.org>.

<sup>9</sup><http://www.imageclef.org/>.

## 2 Plantclef: A 7-Year-Long Evaluation of Image-Based Plant Identification Systems

### 2.1 Methodology

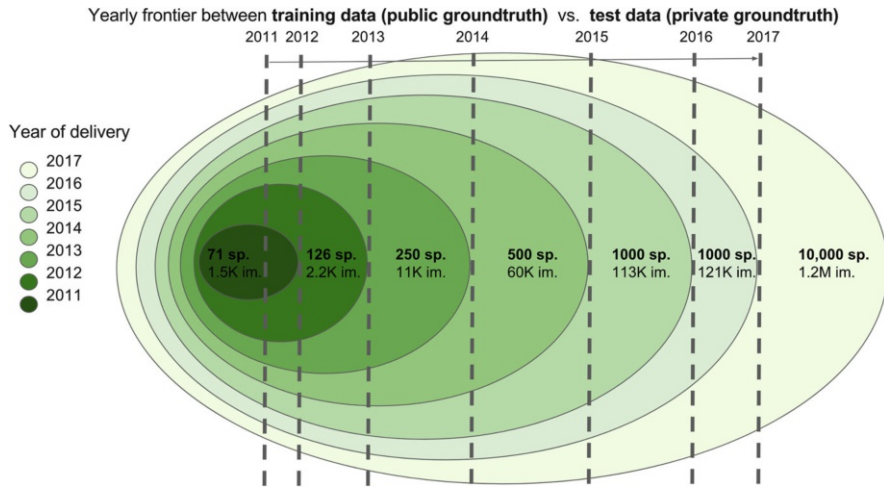
The plant identification challenge of CLEF has been run since 2011, offering today a 7-year follow-up of the progress made in image-based plant identification. A particularity of the benchmark is that it always focused on real-world collaborative data contrary to many other test beds that were created beforehand in the context of well controlled laboratory conditions. Additionally, the evaluation protocol was defined in collaboration with biologists so as to reflect realistic usage scenarios. In particular, we considered the problem of classifying plant observations based on several images of the same individual plant rather than considering a classical image classification task. Indeed, it is usually required to observe several organs of a plant to identify it accurately (e.g. the flower, the leaf, the fruit, the stem, etc.). As a consequence, the same individual plant is often photographed several times by the same observer resulting in contextually similar pictures and/or near-duplicates. To avoid bias, it is crucial to consider such image sets as a single plant observation that should not be split across the training and the test set. In addition to the raw pictures, plant observations are usually associated with contextual and social data. This includes geo-tags or location names, time information, author names, collaborative ratings, vernacular names (common names), picture type tags, etc. Within all PlantCLEF challenges, the use of this additional information was considered as part of the problem because it was judged as potentially useful for a real-world usage scenario.

We provide in Fig. 1 an overview of the data that was shared along the years within the PlantCLEF challenge. Each year, the data was considerably enriched and the number of species was increased from 71 species in 2011 to 10,000 species in 2017 (illustrated by more than 1 million images). This durable scaling-up was made possible thanks to the close collaboration of LifeCLEF with several important actors in the digital botany domain. First of all, the TelaBotanica social network. This network of expert and amateur botanists is one of the largest in the world (with about 40,000 members) and is in charge of many citizen science projects relying on the collection of botanical observations by its members. TelaBotanica develops several collaborative tools dedicated to this purpose, in particular *IdentiPlante*<sup>10</sup> aimed at revising and validating the identification of the observations shared by the network. Most of the data used within the PlantCLEF challenge was collected and revised by the TelaBotanica network. Another source of data were contributions of the users of the PI@ntNet application and the members of the TelaBotanica social network who validated many observations every year.

---

<sup>10</sup><http://www.tela-botanica.org/appli:identiplante> (in French).





**Fig. 1** Overview of the evaluation data used for the PlantCLEF challenge along the years

The evaluation metric that was used from 2011 to 2015 was an extension of the mean reciprocal rank (Voorhees et al. 1999), classically used in information retrieval. The difference is that it is based on a two-stage averaging rather than a flat averaging such as:

$$S = \frac{1}{U} \sum_{u=1}^U \frac{1}{P_u} \sum_{p=1}^{P_u} \frac{1}{r_{u,p}} \tag{1}$$

where  $U$  is the number of image authors within the test set,  $P_u$  the number of individual plants observed by the  $u$ -th author (within the test set),  $r_{u,p}$  is the rank of the correct species within the ranked list of species returned by the evaluated system (for the  $p$ -th observation of the  $u$ -th author). If the correct species does not appear in the returned list, its rank  $r_{u,p}$  is considered as infinite. Overall, the proposed metric makes it possible to compensate the long-tail distribution effects of social data. As in any social network, a few people actually produce huge quantities of data whereas the vast majority of contributors (the long tail) produce much less data.

## 2.2 Main Outcomes

Tables 1 and 2 give a year-to-year overview of the shared data and of the best performing systems (detailed descriptions of the results and systems can be found in the technical overview papers of each year (Goëau et al. 2011b, 2012b, 2013a, 2014, 2015, 2016a, 2017a) and participant working notes papers. To allow a

**Table 1** Three-year synthesis of the PlantCLEF challenge restricted to *leaf scans* and *pseudo-scans*

| Year | #Species | #Images | Evaluated systems | Score of best system | Brief description of best system  |
|------|----------|---------|-------------------|----------------------|---|
| 2011 | 71       | 3967    | 20                | 0.574                | Various local features (around Harris points) + Hash-based indexing + RANSAC based matching |
| 2012 | 126      | 9356    | 30                | 0.565                | Shape and texture global features + SVM classifier  |
| 2013 | 250      | 11,031  | 33                | 0.607                | Shape and texture global features + SVM classifier  |

**Table 2** Seven-year synthesis of the results of the PlantCLEF challenge

| Year | #Species | #Images   | Evaluated systems | Performance of best system | Brief description of best system   |
|------|----------|-----------|-------------------|----------------------------|--|
| 2011 | 71       | 1469      | 20                | 0.251                      | Model-driven segmentation Shape features. Random forest  |
| 2012 | 126      | 2216      | 30                | 0.320                      | Multi-scale local (color) texture SIFT + Sparse coding Spatial pyramidal matching. Linear SVM                            |
| 2013 | 250      | 11,046    | 33                | 0.393                      | Dense-SIFT, C-SIFT, Opponent SIFT HSV-SIF, self-similarity SSIM. Fisher vectors. Linear logistic regression. Late fusion |
| 2014 | 500      | 60,962    | 28                | 0.471                      | ROI segmentation dense-SIFT + Color Moment. Fisher vectors. SVM on FVs   |
| 2015 | 1000     | 113,205   | 18                | 0.667                      | GoogLeNet CNN. Five-fold bagging + Borda fusion  |
| 2016 | 1000     | 121,205   | 29                | 0.827                      | VGGNet, combine outputs of a same observation  |
| 2017 | 10,000   | 1,256,287 | 28                | 0.92                       | Average of many fine-tuned CNNs  |

comprehensive comparison along the years, we isolated in Table 1 the *leaf scans* and *white background* image categories that were part of the evaluation of the three first years but that were abandoned afterwards. Table 2 focuses on photographs of plants in their natural environment (only leaves in 2011–2012, diverse organs and plant views in the following years). For a fair comparison, we also removed from the overview, the submissions that were humanly assisted in some point (e.g. involving a manual segmentation of the leaves).

The main conclusion we can derive from the results of Table 1 is that the classical approach to plant identification consisting of analyzing the morphology of the leaves reached its limit. Leaf shape boundary features and shape matching techniques have been studied for 30 years and can be considered as sufficiently mature for capturing shape information in a robust and invariant way. The limited performance is thus rather due to the intrinsic limitation of using only the leaf morphology for discriminating a large number of species. The fact that scientists focused on leaf-based identification for many years is more related to the fact that the leaf was

easier to scan and to process with state-of-the-art computer vision techniques of that period (segmentation, shape matching, etc.). With the arrival of more advanced computer vision techniques, we were progressively able to make use of other parts of the plant such as flowers or fruits, and to work on larger number of species. For this reason, metrics on leaf scans were abandoned from the PlantCLEF evaluation after 2013.

Table 2 gives the 5-year synthesis of this approach to plant identification that we promoted through PlantCLEF. The most interesting conclusion we can derive is that we observed considerable improvements of the scores along the years whereas the difficulty of the task was increasing. The number of classes almost doubled every year between 2011 and 2015, starting from 71 species in 2011 and reaching 10,000 species in 2017. The increase of the performance can be explained by two major technological breakthroughs.

The first was the use of *aggregation-based* or *coding-based* image representation methods such as the Fisher Vector representation (Sánchez et al. 2013), which was used by the best performing system of Nakayama (2013) and Chen et al. (2014). These methods consist of producing high-dimensional representations of the images by aggregating previously extracted sets of hand-crafted local features into a global vector representation. They rely on a two step process: (1) the learning of a set of latent variables that explain the distribution of the local features in the training set (denoted as the codebook or vocabulary), and (2) the encoding of the relationship between the local features of a given image and the latent variables. Overall, this allows to embed the fine-grained visual content of each image into a single representation space in which classes are easily separable even with linear classifiers.

The second technological step explaining the latest increase of performance is the use of deep learning methods, in particular convolutional neural networks (CNN) such as GoogLeNet (Szegedy et al. 2015). In 2015, the 10 best evaluated systems were based on CNNs. The performance difference is mainly due to particular system design improvements such as the use of bagging in the best run of Choi (2015b). CNNs recently received a high amount of attention caused by the impressive performance they achieved in the ImageNet classification task (Krizhevsky et al. 2012). The force of these technologies relies on their ability to learn discriminant visual features directly from the raw pixels of the images without falling into the trap of the curse of dimensionality. This is achieved by stacking multiple *convolutional layers*, i.e. the core building blocks of a CNN. A convolutional layer basically takes images as input and produces as output *feature maps* corresponding to different convolution kernels, i.e. looking for different visual patterns. Looking at the impressive results achieved by CNN's in the 2015 edition of PlantCLEF there is absolutely no doubt that they are able to capture discriminant visual patterns of the plants in a much more effective way than previously engineered visual features. The editions of PlantCLEF 2016 and 2017 have also clearly confirmed the capacity of CNNs to take advantage of large noisy datasets. Indeed, in the 2017 edition, all networks trained solely on the noisy dataset (coming from web crawl) outperformed the same models trained on the trusted data (coming from the trusted Encyclopedia

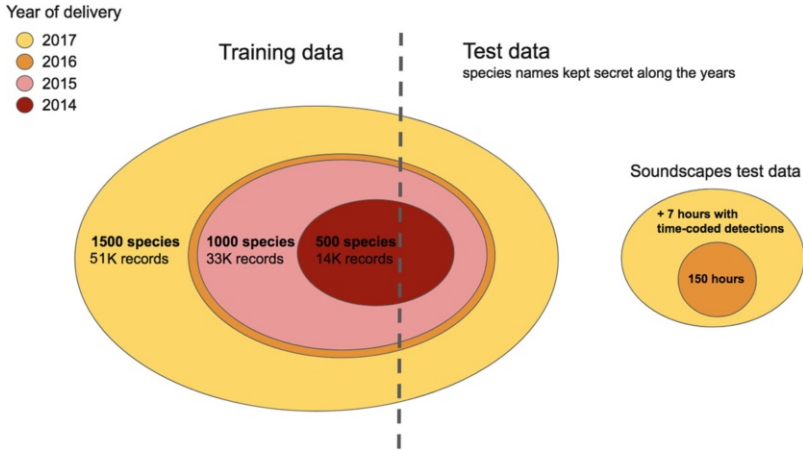
of Life website). Even at a constant number of training iterations (i.e. at a constant number of images passed to the network), it was more profitable to use the noisy training data. This means that diversity in the training data is a key factor to improve the generalization ability of deep learning. The noise itself seems to act as a regularization of the model. The amazing performance of the best runs, which reached a score higher than 90% of correct identification on 10,000 classes opens new perspectives on the potential of automated plant species capacities at the world level.

### 3 Birdclef: A 4 Year Long Evaluation of Bird Sound Identification Systems

#### 3.1 Methodology

The bird identification challenge of LifeCLEF, initiated in 2014 in collaboration with Xeno-Canto, considerably increased the scale of the seminal challenges. The first bird challenge ICML4B (Glotin et al. 2013a) initiated in 2012 by DYNI/SABIOD had only 35 species, but received 400 runs. The next at MLSP had only 15 species, the third (NIPS4B (Glotin et al. 2013b) in 2013 by SABIOD) had 80 species. Meanwhile, Xeno-canto, launched in 2005, hosts bird sounds from all continents and daily receives new recordings from some of the remotest places on Earth. It currently archives with 379,472 recordings, 9779 species of birds, making it one of the most comprehensive collections of bird sound recordings worldwide, and certainly the most comprehensive collection shared under Creative Commons licenses.

For the BirdCLEF challenge, it was decided to not consider the whole Xeno-Canto dataset but to rather focus on a specific region. The objective was to find a good trade-off between scalability and biodiversity coverage. A sufficient number of species had to be considered so as to evaluate the feasibility of a real-world biodiversity monitoring system. But on the other side, it was necessary to limit the volume of data to be processed by the participating research groups so as to mitigate computational challenges and data management. The chosen region of interest has been the Amazonian rain forest because it is one of the richest in the world in terms of biodiversity but also one of the most endangered. For the first edition of the challenge, in 2014, the evaluation dataset was restricted to the 500 species having the most records in an Amazonian area straddling Brazil and neighboring countries. The geographical extent and the number of species were progressively increased over the years so as to reach 1000 species in 2015/2016, and 1500 in 2017. By nature, the Xeno-Canto data as well as the BirdCLEF subset has a massive class imbalance. For instance, the 2017 dataset contains 48,843 recordings in total, with a minimum of four recordings for *Laniocera rufescens* and a maximum of 160 recordings for *Henicorhina leucophrys*.



**Fig. 2** Overview of the evaluation data used for the BirdCLEF challenge along the years

A comprehensive overview of the data shared<sup>11</sup> over the years is provided in Fig. 2. Each year, selected Xeno-canto recordings were split in two parts: 2/3 of the data was shared as training data so as to allow participants to train and optimize their system, and the other 1/3 of the recordings were kept as official test samples and shared to the participants a few weeks after the training set. To avoid participants tuning their system on the test data, the species names were removed from the test set and kept secret over the years (i.e. participants have to run their system in a blind manner). To allow a long-term evaluation of the progress made, it was also ensured that the test data provided each year were a superset of the test data of the previous years. Furthermore, the recordings were shared using a stable format along the years. Each audio file was associated with an XML file containing the available meta-data such as the date, the geo-location, the author, the type of sound (call, song, alarm, flight, etc.) or some collaborative quality ratings. For the training set, the meta-data also included the information related to the species of the bird(s) vocalizing within the recording (taxonomic names and sometimes common names). Most Xeno-Canto recordings are captured using mono-directional devices in order to focus on a single vocalizing bird. The name of the species of this primary singing bird is annotated in the meta-data through a field entitled “foreground species”. But often, there is also a number of other birds that can be heard in the background. The names of the species of the other birds are often annotated in the meta-data through a field entitled “background species”.

Identifying birds from mono-directional recordings such as the ones discussed above is of high interest for many scenarios. In particular, this could help non-experts as well as experts in the process of collecting and identifying such new

<sup>11</sup>Some sample can be listen at <http://sabiod.org/DYNITAG/BIRDCLEF>.

recordings. To complement this, there is also an interest in identifying birds from *omnidirectional* recordings (i.e. the target is the foreground species) or *soundscape*. This enables more passive monitoring scenarios such as setting up a network of static recorders that would continuously capture the surrounding sound environment. Therefore, we started to integrate soundscape recordings within the BirdCLEF challenge in 2016. A significant number of recordings tagged as *soundscapes* actually already existed in the Xeno-Canto collection. They usually correspond to longer recordings than the mono-directional ones and they do not have any *foreground* species in the meta-data. 925 of such soundscapes were found in the Amazonian area and were integrated as a new test within the BirdCLEF 2016 challenge. One of the limitations of this new content, however, was that the vocalizing birds were not localized in the recordings. The set of species audible in the recording was identified in the meta-data but the vocalizing specimens were not localized in time. Thus, to allow a more accurate evaluation, it was decided to introduce new time-coded soundscapes within the BirdCLEF 2017 challenge. In total, 6.5 hours of recordings were collected in the Amazonian forests and were manually annotated by two experts including a native of the Amazon forest, in the form of time-coded segments with associated species name.

The evaluation protocol of BirdCLEF remained roughly the same during the 4 years it ran. Participants were asked to run their system so as to identify all the actively vocalizing bird species in each test recording (or in each test segment of 5 s for the soundscape). Up to 4 *run files* per participant could be submitted to allow evaluating different systems or system configurations (a *run file* is a formatted text file containing the species predictions for all test items). Each species had to be associated with a normalized score in the range [0, 1] reflecting the likelihood that this species is singing in the test sample. For each submitted run, participants had to signal if the run was performed fully automatically or with human assistance, and if they used a method based only on audio analysis or with the use of the metadata. The evaluation metric used was the mean Average Precision (mAP) averaged across all queries, considering each recording in the test set as a query and computed as:

$$mAP = \frac{\sum_{q=1}^Q AveP(q)}{Q},$$

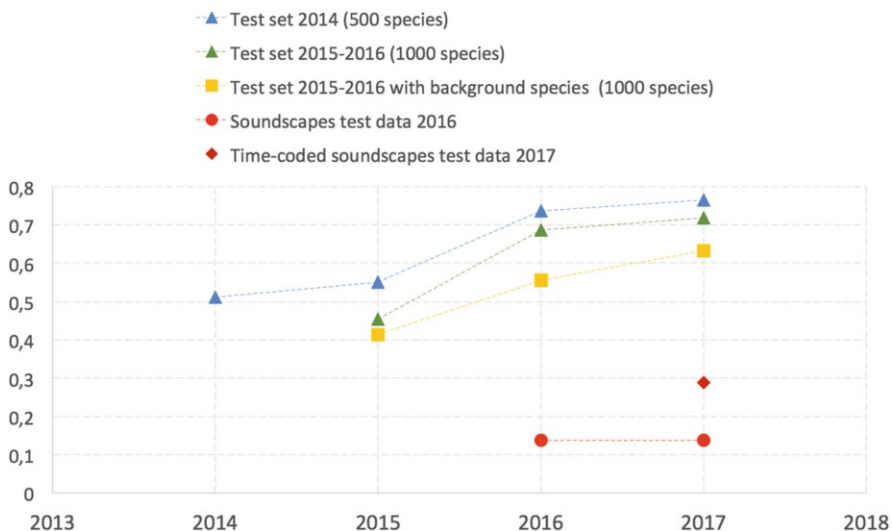
where  $Q$  is the number of test samples and  $AveP(q)$  for a given test file  $q$  is computed as

$$AveP(q) = \frac{\sum_{k=1}^n (P(k) \times rel(k))}{\text{number of relevant documents}}.$$

Here  $k$  is the rank in the sequence of returned species,  $n$  is the total number of returned species,  $P(k)$  is the precision at cut-off  $k$  in the list and  $rel(k)$  is an indicator function equaling 1 if the item at rank  $k$  is a relevant species (i.e. one of the species in the ground truth).

### 3.2 Main Outcomes

Between 60 and 90 research groups registered each year for the BirdCLEF challenge and about 20 of them submitted run files to at least one of the yearly campaigns (with a variation of 5–10 participants depending on the year). This durable evaluation allowed to accelerate the progress made along the years by measuring it accurately thanks to the re-used sub-set test data. As a synthesis of this long-term effort, Fig. 3 displays the evolution of the best mAP scores that were obtained over the years. The curve corresponding to the 2014 test set, in particular, shows the impressive progress that was made from the beginning of the challenge. The best mAP value actually increased from 0.51 to 0.76 in 4 years (for the mono-directional recordings). A big step was particularly observed between 2015 and 2016 (Goëau et al. 2016b). It was exclusively due to the progress of the underlying methods and algorithms since the training set shared within BirdCLEF was exactly the same for these 2 years (as illustrated in Fig. 2). More precisely, and without great surprise, the best system evaluated in 2016 was the first one using deep learning technologies. The convolutional neural network it relied on, outperformed by 4% the mAP of the previous state-of-the-art method of 2014 and 2015, which was based on strong feature engineering and classical machine learning algorithms. After this first remarkable success, most participants in the BirdCLEF challenge continued exploring the use of CNNs in 2017. The different systems used in 2017 mainly differed in the employed CNN architecture and in the time-frequency representation given as input of the CNN. Interestingly, the best system in 2017, from DYNIS



**Fig. 3** Overview of the performance of the best systems evaluated within the BirdCLEF challenge (for different test data sets)

CNRS team (Sevilla and Glotin 2017; Joly et al. 2017), was an adaptation of the Inception model (version 4), i.e. a CNN that was designed by Google for large scale image classification tasks. This raw model was fine-tuned directly from the weights of the initial image classifier. This illustrates the strong convergence of machine learning methods for different contents and the feasibility of transferring knowledge from one modality to another, as long as one uses a common representation (i.e. 2D time-frequency images). The second main outcome of BirdCLEF which can be observed in Fig. 3, is that the soundscape task appears to be much more challenging than the classical task that we shall consider here as mono-species recordings. The identification performance actually remains pretty high for the mono-species recordings, even when considering all the species vocalizing in the background (yellow curve). On the contrary, the best mAP obtained on the 2016 soundscape data set is very low and did not improve between 2016 and 2017 (red curve). One of the main difficulties of such recordings is that many individual birds of several species are often singing simultaneously. This profusion of overlapping sources causes the classical CNN models trained on the mono-species to fail. A good method on the soundscape task seemed to be the feature engineering based method of Lasseck (2015), as the deep learning methods employed by the other participants in 2016 and 2017 were less efficient on the non time-coded soundscape 2016 test set. It is likely that the strength of the features engineering method is based on the extraction of very species-specific time-frequency features. This expert fine-grained approach may allow the extraction of features more robust to the species overlap problem. This verdict was the main reason why we introduced a new soundscape dataset in 2017 (Goëau et al. 2017b), in the form of time-coded segments of 5 s, each associated with the list of species vocalizing in this small segment. The goal was to encourage the participants to output predictions at that temporal resolution instead of processing the whole soundscape as a classical recording. The performance achieved on this new test set (dark red point Fig. 3) confirmed that the temporal resolution of the prediction was one of the issues and that processing each chunk of 5 s separately improves the results over the previous soundscape test set. However, the best performance remains much lower than for the mono-species task. One of the most likely reasons is the bias between the training data (mono-species) and the test data (soundscape). The overlap of all the birds vocalizing simultaneously actually induces audio patterns that cannot be captured directly from the mono-species recordings. A solution to learn such patterns would be to integrate soundscape with time-coded annotations in the training set itself. This approach is unfortunately not realistic because of the cost to produce such content. Another more realistic perspective is to run data augmentation synthesizing new training data from the mono-species recordings themselves. The improvement of the quality of automatic bird activity detection (BAD) is also being taken in consideration as recently depicted in the BAD challenge (Stowell et al. 2016). Finally, we are investigating a more advanced paradigm towards binaural source diarization and joint classification from stereo soundscape in future BirdCLEF sessions.



## **4 SeaCLEF: A 4-Year Evaluation of Sea Organisms Identification**

The need for automated methods for sea-related multimedia data is driven by the recent sprout of marine and ocean observation approaches (mainly imaging systems) and their employment for marine ecosystem analysis and biodiversity monitoring. Indeed in recent years we have witnessed an exponential growth of sea-related multimedia data in the forms of images/videos/sounds, for disparate reasons ranging from fish biodiversity monitoring to marine resource managements to fishery to educational purposes. However, the analysis of such data is particularly expensive for human operators, thus limiting the impact that the technology may have in understanding and sustainably exploiting the sea/ocean. Within LifeCLEF, we investigated several highly demanding annotation scenarios including coral reef fish species monitoring, humpback whale individual recognition, salmon detection for water turbine monitoring and picture-based marine animal species recognition. In the following two subsections, we give an overview of the two challenges that attracted the most participants and that were conducted over several consecutive years.

### ***4.1 Underwater Coral Reef Species Monitoring: Methodology and Main Outcomes***

Underwater imaging systems are increasingly used in a range of monitoring or exploratory applications, in particular for biological (e.g. benthic community structure, habitat classification), fisheries (e.g. stock assessment, species richness), geological (e.g. seabed type, mineral deposits) and physical surveys (e.g. pipelines, cables, oil industry infrastructure). Their usage has benefitted from the increasing miniaturization and cost-effectiveness of submersible ROVs (remotely operated vehicles) and advances in underwater digital cameras. These technologies have revolutionized our ability to capture high-resolution images in challenging aquatic environments and are also greatly improving our ability to effectively manage natural resources, increasing our competitiveness and reducing operational risks in industries that operate in both marine and freshwater systems. Despite these advances, the analysis of the produced data usually requires very time-consuming and expensive input by human observers. This is particularly true for ecological and fishery video data, which often requires laborious visual analysis. This analytic bottleneck greatly restricts the use of these otherwise powerful video technologies and demands effective methods for automatic content analysis to enable proactive provision of analytic information. The underwater video dataset used within LifeCLEF was derived from the Fish4Knowledge video repository, which contains about 700,000 10-min video clips that were taken in the past 5 years to monitor Taiwan's coral reefs. The Taiwan area is particularly interesting for studying the

marine ecosystem, as it holds one of the largest fish biodiversities of the world with more than 3000 different fish species.<sup>12</sup> The dataset contains videos recorded from sunrise to sunset showing several phenomena, e.g. murky water, algae on camera lens, etc., which make the identification task more complex. Each video has a resolution of either  $320 \times 240$  or  $640 \times 480$  with 5–8 fps.

The data set used for the coral reef challenge of LifeCLEF 2015, LifeCLEF 2016 and LifeCLEF 2017 was a small annotated subset of the Fish4Knowledge repository (Spampinato et al. 2016). It was composed of about 90 videos manually annotated for a list of 15 fish species. Each video was labelled and agreed by two expert annotators and the ground truth consists of a set of bounding boxes (one for each instance of the given fish species list) together with the fish species. In total the dataset contained more than 9000 annotations (bounding boxes + species) with a relatively high imbalance in the number of instances of fish species: for instance it contained 3165 instances of “*Dascyllus Reticulates*” and only 72 instances of “*Zebrosoma Scopas*”. For each considered fish species, its fishbase.org link was also given. In the fishbase webpage, participants could find more detailed information about fish species including also high quality images that could be used as additional training data. In order to make the identification process independent from tracking, temporal information was not exploited. This means that the annotators only labelled fish for which the species was clearly identifiable, i.e., if at frame  $t$  the species of fish A is not clear, it was not labelled, no matter if the same fish was in the previous frame ( $t-1$ ). Each video was accompanied by an xml file that contains instances of the provided list species as well as information on the camera location e.g.

```
<?xml version="1.0" encoding="utf-8"?>
<video id="0b21f0579d247c855e05405d3ed805c1#201205251240" location="NPP3" camera="4">
  <frame id="0">
    <object fish_species="Dascyllus Aruanus" h="68" w="87" x="322" y="233"/>
  </frame>
  <frame id="1">
    <object fish_species="Dascyllus Aruanus" h="68" w="87" x="319" y="230"/>
  </frame>
  <frame id="2">
    <object fish_species="Dascyllus Aruanus" h="68" w="87" x="342" y="231"/>
  </frame>
  <frame id="391">
    <object fish_species="Plectrogly-Phidodon Dickii" h="50" w="35" x="271" y="336"/>
    <object fish_species="Plectrogly-Phidodon Dickii" h="41" w="29" x="339" y="375"/>
  </frame>
</video>
```

Since the end-to-end objective of the task was to count the number of specimens per species (for biodiversity monitoring), we introduced two related evaluation

<sup>12</sup>For which a taxonomy is available at <http://fishdb.sinica.edu.tw>.

metrics: the “**Counting Score (CS)**” and the “**Normalized Counting Score (NCS)**”, defined as:

$$CS = e^{-\frac{d}{N_{gt}}} \quad (2)$$

with  $d$  being the difference between the number of occurrences in the run (per species) and,  $N_{gt}$ , the number of occurrences in the ground truth. The Normalized Counting Score instead depends on precision  $Pr$ :

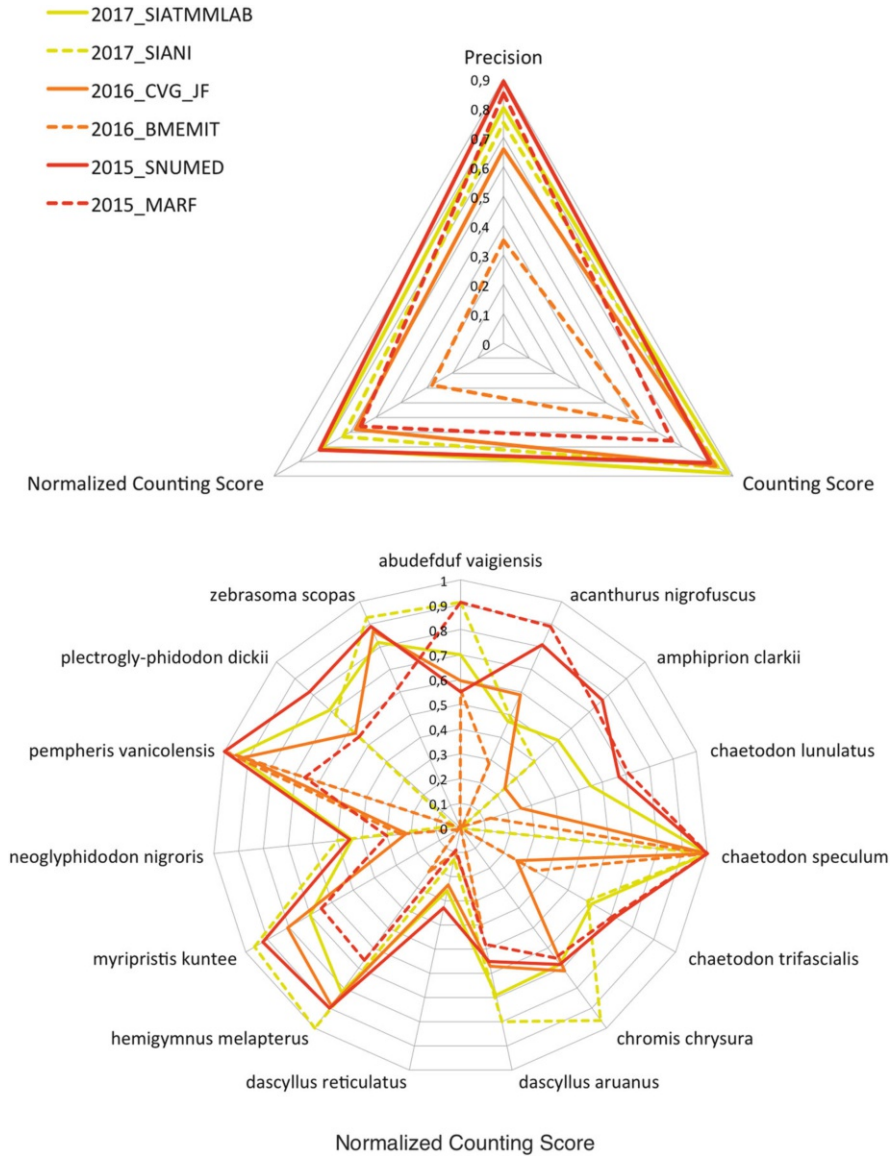
$$NCS = CS \times Pr = CS \times \frac{TP}{TP + FP} \quad (3)$$

with  $Pr = TP/(TP + FP)$ ,  $TP$  and  $FP$  being the true positives and the false positives. A detection was considered as true positive if the intersection over union score of its bounding box and the ground truth was over 0.5 and the species was correctly identified.

Figure 4 shows an overview of the score obtained by the best systems evaluated in 2015, 2016 and 2017 on the same coral reef data set. In the previous fish classification challenge the hierarchical LBP classifier (DYNi team Joalland et al. 2014) won. However, CNN was the best system of 2015 (by SNUMED Choi 2015a), and it was not outperformed in the following years. Contrary to all other LifeCLEF challenges, no real progress were thus observed over the years. The system of SIATMMLAB in 2017 (Zhuang et al. 2017) was devised as an improvement of the one of SNUMED but its precision was still lower resulting in a lower Normalized Counting Score in the end. The right plots of Fig. 4 show that the main strength of the SNUMED system is to be more stable than the other systems across the different species. Importantly, this is rather due to a better detection of the candidate fish instances than a better performance of the classification of the resulting bounding boxes. The SIATMMLAB system actually used a more advanced convolutional neural network model for the classification but it was less accurate in the preliminary detection phase.

## 4.2 Individual Whale Identification: Methodology and Main Outcomes

The problem of automatically identifying individual organisms rather than species has received much less attention. Yet, for some groups, it is preferable to monitor the organisms at the individual level rather than at the species level. This is notably the case of big animals, such as whales and elephants, of which the populations are scarcer and are traveling longer distances. Monitoring individual animals allows gathering valuable information about population sizes, migration, health, sexual maturity and behavior patterns. Tracking devices and tagging technologies are



**Fig. 4** Overview of the performance of the best systems evaluated within the coral reef species recognition challenge

only part of the solution because of their invasive character, relatively high cost and limited lifetime. Morphological/biometric approaches are a complementary approach that is less invasive, more durable and cheaper for nature watchers mobilized on a given spot. Using natural markings to identify individual animals

over time is usually known as photo-identification. This research technique is used on many species of marine mammals. Initially, scientists used artificial tags to identify individual whales, but with limited success (most tagged whales were actually lost or died). In the 1970s, scientists discovered that individuals of many species could be recognized by their natural markings. These scientists began taking photographs of individual animals and comparing these photos against each other to identify individual animal movements and behavior over time. Since its development, photo-identification has proven to be a useful tool for learning about many marine mammal species including humpbacks, right whales, finbacks, killer whales, sperm whales, bottlenose dolphins and other species to a lesser degree. This process is still mostly done manually making it impossible to get an accurate count of all the individuals in a given large collection of observations. Researchers usually survey a portion of the population, and then use statistical formulae to determine population estimates. To limit the variance and bias of such an estimator, it is however required to use sufficiently large samples that still make it a very time-consuming process. Automating the photo-identification process could drastically scale-up such surveys and open brave new research opportunities for the future.

To evaluate this scenario, we did set up a test-bed in collaboration with Cetamada, a Malagasy Non-Profit Association created in May 2009, whose goal is to protect marine mammal population and their habitat in Madagascar through sustainable eco-tourism and scientific research. There are presently four citizen sciences data collection sites (St. Marys, Majunga, Ifaty and Fort Dauphin) for which hotel-establishments and their customers have become sentinels for data collection. This method helps obtain more than 250 photo IDs each year, which effectively helps produce a photo catalogue of humpback whales reproducing on Malagasy coasts. From that data, we built an evaluation dataset of 2005 images of humpback whales that were collected between 2009 and 2014. After acquisition, each photograph was manually cropped so as to focus only on the caudal fin that is the most discriminant pattern for distinguishing one individual whale from another. Actually, the fins can be distinguished thanks to the natural markings and/or the scars that appear along the years. Automatically finding such matches in the whole dataset and rejecting the false alarms is difficult for three main reasons. The first reason is that the number of individuals in the dataset is high, around 1200, so that the proportion of true matches is actually very low (around 0.05% of the total number of potential matches). The second difficulty is that distinct individuals can be very similar at a first glance and that it is often difficult to distinguish them even for a human annotator. To discriminate the true matches from such false positives, it is required to detect very small and fine-grained visual variations such as in a spot-the-difference game. The third difficulty is that all images have a similar water background of which the texture generates quantities of local mismatches.

Concretely, the task consisted in detecting as many true matches as possible from the whole dataset, in a fully unsupervised way. Each evaluated system had to return

a *run file* (i.e., a raw text file) containing as many lines as the number of discovered matches, each match being a triplet of the form:

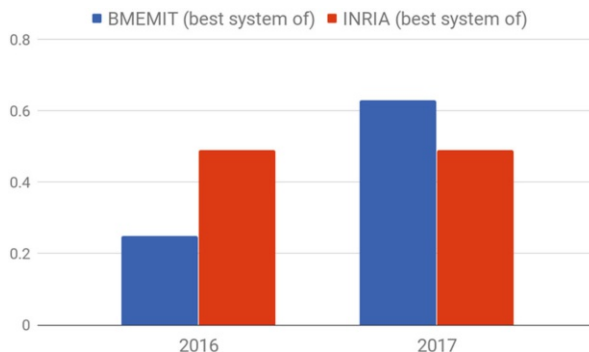
$$\langle \text{imageX.jpg imageY.jpg score} \rangle$$

where *score* is a confidence score in  $[0, 1]$  (1 for highly confident matches). The retrieved matches had to be sorted by decreasing confidence score. A run should not contain any duplicate match (e.g.,  $\langle \text{image1.jpg image2.jpg score} \rangle$  and  $\langle \text{image2.jpg image1.jpg score} \rangle$  should not appear in the same run file). The metric used to evaluate each run was Average Precision:

$$\text{AveP} = \frac{\sum_{k=1}^K P(k) \times \text{rel}(k)}{M}$$

where  $M$  is the total number of true matches in the groundtruth,  $k$  is the rank in the sequence of returned matches,  $K$  is the number of retrieved matches,  $P(k)$  is the precision at cut-off  $k$  in the list, and  $\text{rel}(k)$  is an indicator function equaling 1 if the match at rank  $k$  is a relevant match, 0 otherwise.

The same challenge was run for two consecutive years, in 2016 and 2017. An overview of the results achieved by the best system of each participant (yearly) is provided in Fig. 5. In 2016, the best result was achieved by the INRIA-ZENITH team who used a large-scale matching system based on SIFT visual features, approximate k-nn search and a RANSAC-like spatial consistency refinement step to reject false positives (Joly et al. 2016). In 2017, a similar system was re-implemented by BMEMIT and extended with an additional clustering step which provided a consistent improvement (Dávid Papp and Szűcs 2017). Interestingly, the whale photo-identification challenge is the only one within LifeCLEF for which deep learning technologies do not provide the best performance (although several attempts were made). The main reason is that it is very different from the



**Fig. 5** Overview of the performance of the best systems evaluated within the whale photo-identification challenge

classical challenges studied in the machine learning community. This is actually an unsupervised classification problem but for which the visual patterns to be discovered are very small and lost among a high amount of other highly similar patterns. Only an explicit spatial verification based on the hypothesis of an epipolar geometry allows to distinguish the real matches from the distractors. Without supervision, convolutional neural networks fail to capture this property.

## 5 Cross-Task Analysis of the Use of Contextual Meta-Data

Most of the data sets shared within LifeCLEF since 2011 included contextual meta-data in addition to the raw audio-visual contents. As an illustration, Table 3 lists the meta-data shared for each image of the training set of PlantCLEF 2016. A large fraction of the plant and bird observations, in particular, were associated with their date and geo-location. This information was expected to be highly useful for species identification. Indeed, most plants and animals live in specific ecological niches and are likely to be observed at some specific periods.

Table 4 reports the results obtained by the participants of the plant and the bird tasks who attempted to evaluate the potential benefit of this meta-data over the years. However, the benefit of using the temporal and spatial information has never been decisive in any of the LifeCLEF challenges. Worse, it often degraded the performance compared to using the raw audio-visual data solely. To better highlight

**Table 3** Types of metadata shared within PlantCLEF challenge

| Type of metadata     | Metadata description  |
|----------------------|---|
| Observation Id       | The plant observation ID from which several pictures can be associated  |
| Media Id             | The ID of the image   |
| View content         | Description of the content visible in the image :<br>Entire or branch or flower or fruit or leaf or leafScan, etc.  |
| Class Id             | The class number ID that must be used as ground-truth.<br>It is a numerical taxonomical number used by Tela Botanica  |
| Species name         | The species names (containing three parts: the genus name,<br>the specific epithet, the author(s) who discovered or<br>revised the name of the species)   |
| Family               | The name of the family, two levels above the species in<br>the taxonomical hierarchy used by Tela Botanica  |
| Date                 | (If available) the date when the plant was observed   |
| Vote                 | The (round up) average of the user ratings of image quality   |
| Location             | (If available) locality name, a town most of the time   |
| Latitude & longitude | (If available) the GPS coordinates of the observation in the EXIF metadata,<br>or if no GPS information were found in the EXIF the GPS coordinates<br>of the locality where the plant was observed<br>(only for the towns of metropolitan France) |
| Author               | Name of the author of the picture   |
| YearInCLEF           | ImageCLEF2011, ImageCLEF2012, ImageCLEF2013,<br>PlantCLEF2014, PlantCLEF2015  |

**Table 4** Impact of the use of metadata for plants and birds identification

| Year | Task      | Team                                     | Metadata type        | Improvement |
|------|-----------|--|----------------------|-------------|
| 2011 | PlantCLEF | UAIC                                     | GPS, Date, Author Id | -35.89%     |
| 2012 | PlantCLEF | BTU DBIS<br>(Böttcher et al. 2012)       | GPS                  | -4.76%      |
| 2013 | PlantCLEF | Inria (Bakic et al. 2013)                | Date                 | +9.06%      |
| 2015 | PlantCLEF | SABANCI-OKAN<br>(Ghazi and Ozdemir 2015) | Date                 | +1.23%      |
| 2014 | BirdCLEF  | Inria<br>(Joly et al. 2014a)             | GPS, Date            | +11.28%     |
| 2017 | BirdCLEF  | TUCMI<br>(Kahl et al. 2017)              | GPS                  | -32.67%     |

this finding, Table 4 provides an overview of all the experiments for which it was possible to evaluate the performance of the same system with or without the use of meta-data. The best improvement was achieved by the Inria team in 2013 for the plant task and 2014 for the bird task. Both were obtained by post-filtering the list of candidate species based on a temporal histogram constructed for each species based on the training meta-data. However, these runs were still outperformed by purely content-based methods developed by other participants.

This difficulty of successfully using geography and seasonality is quite surprising. It is actually accepted that the habitat of a given species is highly correlated with its ecological profile. Several reasons explain this paradox. The first one is that the occurrence data of the training set is too sparse to accurately model the distribution of the species. The second reason is that the used machine learning techniques were too straightforward to well address the problem. As discussed in Sect. 6, species distribution modeling from occurrence data is still a hard problem in ecology, in particular in the context of uncontrolled observations such as the one used in the PlantCLEF challenge.

Concerning the use of the observation date, which was the second most used meta-data by participants, there are several difficulties to appropriately exploit it. First, the plant phenology (plant life cycle events) for a given species is different according to its location (i.e. the same species will present different flowering periods, if individuals are not at the same altitude in mountain conditions, are not exposed along the year to the same light conditions, etc.). Secondly, it's now well accepted that the plant phenology for a given species is changing from 1 year to another one, according to the climate changes. It is then difficult to find a regular pattern over several years, even if observations are produced at the same location. Thirdly, as plant phenology is profoundly influenced by human activity (fertilizer, pruning, greenhouse cultivation, etc.), the phenology of most of the plants observed in urban areas can be different than the individuals growing in natural conditions. According to these various factors, and the limited number of observations per species, one can understand that it is not easy to find a method which is robust on a large scale



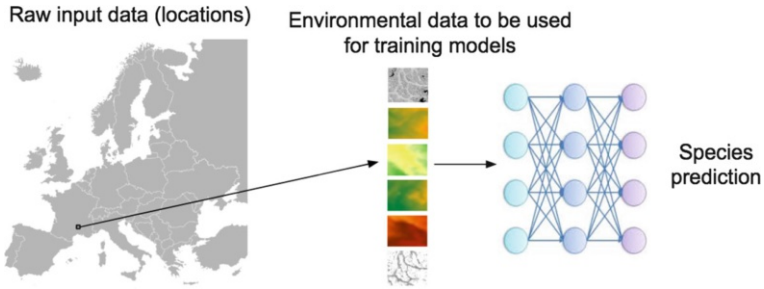
for a strong improvement of the identification performance. The potential of the use of meta-data, which is recognized as highly relevant by naturalists, has still to be demonstrated, and will be a central part of a new challenge entitled *GeoLifeClef*, that will be launched in 2018.

## 6 GeoLifeClef: A Machine Learning Approach to Species Distribution Modeling

In order to increase the interest of the computer science community in the use of the *undisclosed* potential of meta-data for automated species identification, we designed a new challenge within LifeCLEF to be ran in 2018 for the first time. In particular, the new task called *GeoLifeClef* will focus on *location-based species recommendation*. Automatically predicting the list of species that are the most likely to be observed at a given location is useful for many scenarios in biodiversity informatics: (a) it could improve species identification processes and tools by reducing the list of candidate species that are observable at a given location; (b) it could facilitate biodiversity inventories through the development of location-based recommendation services (typically on mobile phones) as well as the involvement of non-expert nature observers; (c) last but not least, it might serve educational purposes thanks to biodiversity discovery applications providing functionalities such as contextualized educational pathways. This new challenge will contribute to increase exchanges between the computer science community and ecological statisticians working on species distribution modelling problems, who would both have lots to gain by sharing their experiences and knowledge.

Concretely, the objective of the challenge will be to predict the list of species that are the most likely to be observed at a given location. Therefore a large training set of species occurrences will be provided, each occurrence being associated with a multi-channel image characterizing the local environment. Indeed, it is usually difficult to learn a species distribution model directly from spatial positions because of the limited number of occurrences and the sampling bias. What is usually done in ecology is to predict the distribution on the basis of a representation in the environmental space, typically a feature vector composed of climatic variables (average temperature at that location, precipitation, etc.) and other variables such as soil type, land cover, distance to water, etc. As illustrated in Fig. 6 the originality of GeoLifeCLEF is to generalize such a niche modeling approach to the use of an image-based environmental representation space. Instead of learning a model from environmental feature vectors, the goal of the task will be to learn a model from  $k$ -dimensional image patches, each patch representing the value of an environmental variable in the neighborhood of the occurrence. From a machine learning point of view, the challenge will thus be treatable as an image classification task.

According to the huge volume of new data produced by large scale citizen science initiatives, such as eBird, iNaturalist, or PI@ntNet, and the accessibility of various



**Fig. 6** Overview of the GeoLifeClef challenge

environmental data based on the open science movement, the adaptation potential (to various living organism groups, environments, regions, etc.) of the result of this task is extremely important. The hope with this new task is to open new interdisciplinary research opportunities based on the analysis of a very large amount of data that was never mobilized beforehand.

## 7 Conclusion

This chapter has discussed the experience of running the LifeCLEF challenges from 2011 to 2017. Several large-scale and repeatable experiments were designed over the years in order to boost research on biodiversity information retrieval. A high number of research groups participated in and benefited from this joint research effort. Overall, LifeCLEF has had an important impact in different fields including multimedia information retrieval, machine learning and biodiversity informatics (more than 500 citations at the end of 2017 according to Google scholar). The main lessons we learned in the design of attractive, sustainable and impacting challenges are the following:

- **Data is a key factor:** sharing original, valuable and large-scale data sets is a key factor for attracting researchers on a given challenge. Within LifeCLEF, tens of men months have been spent in integrating, cleaning and annotating the raw content of data providers.
- **Hard problems but simple tasks:** if the task is too specific or too complex in terms of objectives, it is not attractive. For instance, it is crucial to avoid fragmenting the challenge in many subtasks even if at a first glance it can appear as a good way to better understand the results. What happens in practice is that the participation is fragmented as well: only a few systems are run for each subtask and there is not enough output data to conduct relevant analyses. A single task relying on a hard scientific problem is the best way to federate a community around a given topic.

- **Sustaining the community requires a good trade-off between novelty and continuity:** research relies on long-term efforts and investigations. Thus, it is important to avoid switching to a new problem when the previous one is not solved. On the other hand, sticking exactly to the same challenge over years is counterproductive in terms of attractiveness and emulation. The good trade-off consists in progressively increasing the complexity and/or the difficulty of the task but preserving a sufficient continuity to allow former participants to build on top of their acquired knowledge and technologies.

**Acknowledgements** The organization of the PlantCLEF task is supported by the French project Floris’Tic (Tela Botanica, INRIA, CIRAD, INRA, IRD) funded in the context of the national investment program PIA. The organization of the BirdCLEF task is supported by the Xeno-Canto foundation for nature sounds as well as the French CNRS project SABIOD.ORG and EADM MADICS, and Floris’Tic. The annotations of some soundscapes were prepared with the late wonderful Lucio Pando at Explorama Lodges, with the support of Pam Bucur, Marie Trone and H. Glotin. The organization of the SeaCLEF task is supported by the Ceta-mada NGO and the French project Floris’Tic.

## References

- Baillie J, Hilton-Taylor C, Stuart SN (2004) 2004 IUCN red list of threatened species: a global species assessment. IUCN, Gland
- Bakic V, Mouine S, Ouertani-Litayem S, Verroust-Blondet A, Yahiaoui I, Goëau H, Joly A (2013) Inria’s participation at imageclef 2013 plant identification task. In: CLEF (online working notes/labs/workshop)
- Böttcher T, Schmidt C, Zellhöfer D, Schmitt I (2012) Btu dbis’ plant identification runs at imageclef 2012. In: CLEF (online working notes/labs/workshop)
- Cai J, Ee D, Pham B, Roe P, Zhang J (2007) Sensor network for the monitoring of ecosystem: bird species recognition. In: 3rd International conference on intelligent sensors, sensor networks and information, 2007. ISSNIP 2007. <https://doi.org/10.1109/ISSNIP.2007.4496859>
- Chen Q, Abedini M, Garnavi R, Liang X (2014) Ibm research australia at lifeclef2014: plant identification task. In: Working notes of CLEF 2014 conference
- Choi S (2015a) Fish identification in underwater video with deep convolutional neural network: snumedinfo at lifeclef fish task 2015. In: CLEF (working notes)
- Choi S (2015b) Plant identification with deep convolutional neural network: snumedinfo at lifeclef plant identification task 2015. In: Working notes of CLEF 2015 conference
- Dávid Papp FM, Szűcs G (2017) Image matching for individual recognition with sift, ransac and mcl. In: Working notes of CLEF 2017 (cross language evaluation forum)
- Gaston KJ, O’Neill MA (2004) Automated species identification: why not? *Philos Trans Roy Soc Lond B: Biol Sci* 359(1444):655–667
- Ghazi EAOM Berrin Yanikoglu, Ozdemir MC (2015) Sabanci-okan system in lifeclef 2015 plant identification competition. In: Working notes of CLEF 2015 conference
- Glotin H, Clark C, LeCun Y, Dugan P, Halkias X, Sueur J (2013a) Proceedings of 1st workshop on machine learning for bioacoustics - ICML4B. ICML, Atlanta. [http://sabiod.org/ICML4B2013\\_book.pdf](http://sabiod.org/ICML4B2013_book.pdf)
- Glotin H, LeCun Y, Artières T, Mallat S, Tchernichovski HX O (2013b) Proceedings of neural information processing scaled for bioacoustics, from neurons to big data. NIPS International Conference, Tahoe. <http://sabiod.org/nips4b>

- Goëau H, Bonnet P, Joly A, Boujemaa N, Barthélémy D, Molino JF, Birnbaum P, Mouysset E, Picard M (2011a) The imageclef 2011 plant images classification task. In: CLEF 2011
- Goëau H, Bonnet P, Joly A, Boujemaa N, Barthelemy D, Molino JF, Birnbaum P, Mouysset E, Picard M (2011b) The CLEF 2011 plant images classification task. In: Petras V, Forner P, Clough P, Ferro N (eds) CLEF 2011 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1177/>
- Goëau H, Bonnet P, Joly A, Yahiaoui I, Barthélémy D, Boujemaa N, Molino JF (2012a) Imageclef2012 plant images identification task. In: CLEF 2012, Rome
- Goëau H, Bonnet P, Joly A, Yahiaoui I, Barthelemy D, Boujemaa N, Molino JF (2012b) The ImageCLEF 2012 plant identification task. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Goëau H, Bonnet P, Joly A, Bakic V, Barthélémy D, Boujemaa N, Molino JF (2013a) The imageclef 2013 plant identification task. In: CLEF, Valencia
- Goëau H, Joly A, Bonnet P, Bakic V, Barthélémy D, Boujemaa N, Molino JF (2013b) The imageclef plant identification task 2013. In: Proceedings of the 2nd ACM international workshop on multimedia analysis for ecological data. ACM, New York, pp 23–28
- Goëau H, Joly A, Bonnet P, Selmi S, Molino JF, Barthélémy D, Boujemaa N (2014) The lifeclef 2014 plant images identification task. In: CLEF, Sheffield
- Goëau H, Bonnet P, Joly A (2015) The lifeclef 2015 plant images identification task. In: CLEF, Toulouse
- Goëau H, Bonnet P, Joly A (2016a) The lifeclef plant identification task 2016. In: CEUR-WS (ed) CLEF, Evora. CLEF2016 working notes
- Goëau H, Glotin H, Vellinga W, Planqué R, Joly A (2016b) Lifeclef bird identification task 2016: the arrival of deep learning. In: Working notes of CLEF 2016 - conference and labs of the evaluation forum, Évora, 5–8 September, 2016, pp 440–449. <http://ceur-ws.org/Vol-1609/16090440.pdf>
- Goëau H, Bonnet P, Joly A (2017a) Plant identification based on noisy web data: the amazing performance of deep learning (lifeclef 2017). CLEF working notes 2017
- Goëau H, Glotin H, Vellinga W, Planqué B, Joly A (2017b) Lifeclef bird identification task 2017. In: Working notes of CLEF 2017 - conference and labs of the evaluation forum, Dublin, Ireland, September 11–14, 2017. [http://ceur-ws.org/Vol-1866/invited\\_paper\\_8.pdf](http://ceur-ws.org/Vol-1866/invited_paper_8.pdf)
- Joalland P, Paris S, Glotin H (2014) Efficient instance-based fish species visual identification by global representation. In: Working notes for CLEF 2014 conference, Sheffield, September 15–18, 2014, pp 785–789. URL <http://ceur-ws.org/Vol-1180/CLEF2014wn-Life-JoallandEt2014.pdf>
- Joly A, Champ J, Buisson O (2014a) Instance-based bird species identification with indiscriminant features pruning - lifeclef2014. In: Working notes of CLEF 2014 conference
- Joly A, Goëau H, Bonnet P, Bakić V, Barbe J, Selmi S, Yahiaoui I, Carré J, Mouysset E, Molino JF, et al (2014b) Interactive plant identification based on social image data. *Ecol Inf* 23:22–34
- Joly A, Lombardo JC, Champ J, Saloma A (2016) Unsupervised individual whales identification: spot the difference in the ocean. In: Working notes of CLEF 2016 (cross language evaluation forum)
- Joly A, Goëau H, Glotin H, Spampinato C, Bonnet P, Vellinga WP, Lombardo JC, Planqué R, Palazzo S, Müller H (2017) Lifeclef 2017 lab overview: multimedia species identification challenges. In: International conference of the cross-language evaluation forum for European languages. Springer, Berlin, pp 255–274
- Kahl S, Wilhelm-Stein T, Hussein H, Klinck H, Kowerko D, Ritter M, Eibl M (2017) Large-scale bird sound classification using convolutional neural networks. In: CLEF 2017
- Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
- Kumar N, Belhumeur PN, Biswas A, Jacobs DW, Kress WJ, Lopez IC, Soares JVB (2012) Leafsnap: a computer vision system for automatic plant species identification. In: European conference on computer vision, pp 502–516

- Lasseck M (2015) Towards automatic large-scale identification of birds in audio recordings. In: Mothe J, Savoy J, Kamps J, Pinel-Sauvagnat K, Jones GJF, SanJuan E, Cappellato L, Ferro N (eds) *Experimental IR meets multilinguality, multimodality, and interaction. Proceedings of the sixth international conference of the CLEF association (CLEF 2015)*. Lecture notes in computer science (LNCS) 9283. Springer, Heidelberg, pp 364–375
- Lee DJ, Schoenberger RB, Shiozawa D, Xu X, Zhan P (2004) Contour matching for a fish recognition and migration-monitoring system. In: *Optics east, international society for optics and photonics*, pp 37–48
- Nakayama H (2013) Nlab-utokyo at imageclef 2013 plant identification task. In: *CLEF 2013*
- Nilsback ME, Zisserman A (2008) Automated flower classification over a large number of classes. In: *Proceedings of the indian conference on computer vision, graphics and image processing*
- Sánchez J, Perronnin F, Mensink T, Verbeek J (2013) Image classification with the fisher vector: Theory and practice. *Int J Comput Vis* 105(3):222–245
- Sevilla A, Glotin H (2017) Audio bird classification with inception-v4 extended with time-frequency attention mechanisms. In: *Working notes CLEF 2017, conference of the evaluation forum, Dublin*. [http://ceur-ws.org/Vol-1866/paper\\_177.pdf](http://ceur-ws.org/Vol-1866/paper_177.pdf)
- Silvertown J, Harvey M, Greenwood R, Dodd M, Rosewell J, Rebelo T, Ansine J, McConway K (2015) Crowdsourcing the identification of organisms: a case-study of ispot. *ZooKeys* (480):125
- Spampinato C, Palazzo S, Joalland P, Paris S, Glotin H, Blanc K, Lingrand D, Precioso F (2016) Fine-grained object recognition in underwater visual data. *Multimedia Tools Appl* 75(3):1701–1720. <https://doi.org/10.1007/s11042-015-2601-x>
- Stowell D, Wood M, Stylianou Y, Glotin H (2016) Bird detection in audio: a survey and a challenge. In: *26th IEEE international workshop on machine learning for signal proceedings, MLSP*, pp 1–6. <https://doi.org/10.1109/MLSP.2016.7738875>
- Sullivan BL, Aycrigg JL, Barry JH, Bonney RE, Bruns N, Cooper CB, Damoulas T, Dhondt AA, Dieterich T, Farnsworth A, et al (2014) The ebird enterprise: an integrated approach to development and application of citizen science. *Biol Conserv* 169:31–40
- Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 1–9
- Towsey M, Planitz B, Nantes A, Wimmer J, Roe P (2012) A toolbox for animal call recognition. *Bioacoustics* 21(2):107–125
- Trifa VM, Kirschel AN, Taylor CE, Vallejo EE (2008) Automated species recognition of antbirds in a mexican rainforest using hidden markov models. *J Acoust Soc Am* 123:2424
- Voorhees EM, et al (1999) The trec-8 question answering track report. In: *Trec*, vol 99, pp 77–82
- Zhuang P, Xing L, Liu Y, Guo S, Qiao Y (2017) Marine animal detection and recognition with advanced deep learning models. In: *Working notes of CLEF 2017*

# From XML Retrieval to Semantic Search and Beyond



## The INEX, SBS, and MC2 Labs of CLEF 2012–2018

**Jaap Kamps, Marijn Koolen, Shlomo Geva, Ralf Schenkel, Eric SanJuan, and Toine Bogers**

**Abstract** INEX ran as an independent evaluation forum for 10 years before it teamed up with CLEF in 2012. Even before 2012 there was considerable collaboration between INEX and CLEF, and these collaborations increased in intensity when CLEF moved beyond its traditional cross-lingual focus in 2009/2010 shifting to include all experimental IR. This led to the merger of CLEF and INEX, and effectively to the inclusion of INEX as a large track or lab into CLEF in 2012. This chapter details the efforts of the INEX lab in CLEF (2012–2014), as well as the ongoing activities as separate labs, under the labels Social Book Search (2015–2016), and Microblog Contextualization (2016–2018).

---

J. Kamps (✉)  
University of Amsterdam, Amsterdam, Netherlands  
e-mail: [kamps@uva.nl](mailto:kamps@uva.nl)

M. Koolen  
Royal Netherlands Academy of Arts and Sciences, Amsterdam, Netherlands  
e-mail: [marijn.koolen@di.huc.knaw.nl](mailto:marijn.koolen@di.huc.knaw.nl)

S. Geva  
QUT, Brisbane, QLD, Australia  
e-mail: [s.geva@qut.edu.au](mailto:s.geva@qut.edu.au)

R. Schenkel  
University of Trier, Trier, Germany  
e-mail: [schenkel@uni-trier.de](mailto:schenkel@uni-trier.de)

E. SanJuan  
University of Avignon, Avignon, France  
e-mail: [eric.sanjuan@uni-avignon.fr](mailto:eric.sanjuan@uni-avignon.fr)

T. Bogers  
Aalborg University Copenhagen, Copenhagen, Denmark  
e-mail: [toine@hum.aau.dk](mailto:toine@hum.aau.dk)

## 1 Introduction

This chapter documents the INEX lab of CLEF, running 2012–2014, including INEX tracks that continued as independent CLEF labs: the CLEF Social Book Search Lab (2015–2016) and the CLEF Cultural Microblog Contextualization Workshop and Lab (2016–2018). The emphasis is on the INEX and Social Book Search Labs as the Microblog Contextualization Lab was still ongoing at the time of writing.

No single chapter can do justice to the massive amount of work done in the INEX and follow up labs, so this chapter is merely meant as a starting point with references to the respective overview papers providing full details on the tracks and resulting test collections. By providing a high level, but comprehensive overview of the wealth of activities spanning many years, we hope to shed some light on important developments in the field during these years, and relations between various activities inside and outside CLEF that may not be immediately apparent for those familiar with only part of the activities. This highlights also two of the key strengths and contributions of INEX in particular, and CLEF in general. First, it is about the people: the open format and volunteer run activities allowed many (young) researchers to get involved in every aspects, from participant discussion to task or track organization, and has educated a generation of researchers now taking up leadership positions in the field. Second, many of the activities had some degree of success inside the respective CLEF campaign, but also have much greater impact by reuse of the benchmarks in publications, including creative reuse in unexpected settings, and by instigating new activities, both happening in future editions of the track, but also outside the lab or outside CLEF.

This chapter is structured as follows. You are now at the end of the introduction in Sect. 1. Next, in Sect. 2, we document the context of INEX joining CLEF, covering the years 2002–2011. The main part of this chapter is in Sect. 3, in which we document the INEX tracks as part of CLEF 2012–2014. This is followed by a discussion of the two follow up labs seeded from INEX: the *Social Book Search* lab as part of CLEF 2014–2016 in Sect. 4, and the *Cultural Microblog Contextualization* workshop and lab as part of CLEF 2016–2018 in Sect. 5. We close off by providing further discussion and reflection in the final Sect. 6.

## 2 INEX Before CLEF

In this section, we will discuss the context of INEX joining CLEF, very briefly covering the years 2002–2011.

## 2.1 INEX 2002–2011

The *INitiative for the Evaluation of XML Retrieval (INEX)* was founded in 2002 by a group of people led by Mounia Lalmas and Norbert Fuhr. Collaborations between INEX and *Conference and Labs of the Evaluation Forum (CLEF)* date back to these early days. For a good overview of the first 10 years of INEX, we refer to the proceedings and overview papers, and to the chapters in the Springer Encyclopedia of Database Systems (Kazai 2009, 2018). We restrict our attention here to the relation between the INEX and CLEF evaluation forums before the merge in 2012.

Perhaps not known by many people, the only source of funding ever supporting INEX was a modest contribution from the DELOS Network of Excellence, an EU funded initiative to support and promote digital libraries in the period 1997–2007 (Thanos and Casarosa 2017). And apart from incidental, large, support from dedicated EU projects, this same, small support from DELOS presented the life-line of CLEF throughout these years. Within DELOS, INEX and CLEF were co-responsible for the evaluation activities, and INEX always promoted the inclusion of CLEF in these activities even when DELOS seemed more keen on the relevance of INEX for the DL use case. DELOS was also heavily involved in the *European Conference on Digital Libraries (ECDL)* conference, now continued as the *Theory and Practice of Digital Libraries (TPDL)* conference, that provided a home for CLEF as one of its satellite workshops. Looking back over the DELOS years, Thanos and Casarosa (2017, p. 305) list both CLEF and INEX as the most important spin-offs of the DELOS activities.

Until CLEF 2010, when the CLEF focus was firmly on multi-language retrieval, there was incidental direct collaboration between INEX and CLEF. Let us just mention two of the main examples. First, as INEX started running a large-scale Wikipedia image retrieval task (2006–2007), but had few multimedia retrieval participants, it was decided to move this task to CLEF and make it part of the booming CLEF ImageCLEF track, which turned out to be a very good choice attracting substantial participation to both the WikipediaMM task (continuing 2008–2011) and to ImageCLEF. Second, INEX was worried about the state of IR evaluation when the scale of test collections increased and the retrieval tasks became more specialized. This led to a SIGIR workshop on the Future of Information Retrieval Evaluation run at SIGIR 2009 in Boston (Kamps et al. 2009), which was jointly organized by the main chairs of INEX, CLEF, TREC, and NTCIR. INEX chairs who took the initiative for this workshop were particularly overwhelmed by the support and encouragement they received from CLEF. This was an important workshop that had considerable impact in years to come, by creating a vision and longer term agenda, and the clear conviction that more direct collaboration between the different evaluation forums was urgently needed.



## 2.2 *INEX Joining CLEF*

In light of the above, there are three major factors that caused the merger of INEX and CLEF in 2012. First, CLEF decided in 2009 to refocus and break free from the earlier multi-language niche, as well as decided to host their own conference as they had simply outgrown the setup as workshop associated with ECDL. As a case in point, CLEF 2009 was far larger than ECDL 2009, even though CLEF was officially still just one of the ECDL satellite workshops. Second, INEX lost its old home of Dagstuhl after 2008, and found it increasingly difficult to organize its own event in December, competing with ADCS, NTCIR, AIRS, . . . , and the holiday season, and was considering moving to a different location and time-slot. Third, with the need for further collaboration between evaluation initiatives in the air, the wish of CLEF to broaden its tracks beyond the traditional multi-language tasks, and the desire of INEX to find a new home, the merger of INEX and CLEF seemed a win-win for everyone involved.

Hence in 2012, INEX and CLEF merged into a single event, and INEX joined as “superlab” with many tracks, similar to ImageCLEF at the time. In 2011, INEX already changed its track structure in anticipation of the CLEF merger. Note that the INEX 2011 workshop was held in December 2011 near Saarbrücken, and the next edition of INEX as part of CLEF was scheduled for September 2012 in Rome. This made INEX 2012 a proverbial nine-months cycle, putting considerable stress on the INEX volunteer-run organization to deliver the baby in time for CLEF 2012.

The core of INEX had always been its *Ad hoc search* track, running keyword and structured queries against a structured corpus, allowing to retrieve any document part as a result, but evaluated as a proper IR or search task against topical relevance. The Ad hoc track also consumed most of the resources, and had the largest number of participants, effectively competing with an increasing number of other tracks. To promote the other activities, the general chairs, in consultation with the INEX steering committee, made the difficult decision to stop the ad hoc track in 2011, and focus on five activities. The last pre-CLEF edition INEX 2011 (Bellot et al. 2012b) featured the following tracks:

**Books and Social Search Track** *investigating techniques to support users in searching and navigating books, metadata and complementary social media. The Social Search for Best Books Task studies the relative value of authoritative metadata and user-generated content using a collection based on data from Amazon and LibraryThing. The Prove It Task asks for pages confirming or refuting a factual statement, using a corpus of the full texts of 50k digitized books.*

**Data Centric Track** *investigating retrieval over a strongly structured collection of documents based on IMDb. The Ad Hoc Search Task has informational requests to be answered by the entities in IMDb (movies, actors, directors, etc.). The Faceted Search Task asks for a restricted list of facets and facet-values that will optimally guide the searcher toward relevant information.*

**Question Answering Track** *investigating tweet contextualization, answering questions of the form “what is this tweet about?” with a synthetic summary of contextual information grasped from Wikipedia and evaluated by both the relevant text retrieved, and the “last point of interest.”*

**Relevance Feedback Track** *investigating the utility of incremental passage level relevance feedback by simulating a searcher’s interaction. An unconventional evaluation track where submissions are executable computer programs rather than search results.*

**Snippet Retrieval Track** *investigating how to generate informative snippets for search results. Such snippets should provide sufficient information to allow the user to determine the relevance of each document, without needing to view the document itself.*

Where as the INEX Relevance Feedback and Snippet Retrieval tracks were mature and continued from earlier years, the other three tracks were completely redesigned in anticipation of the CLEF merger: the social data of the INEX Book Track was added, the tweet contextualization focus emerged in the INEX QA Track, and work on highly structured data was started at the INEX Data Centric Track.

### 3 INEX at CLEF

In this section, we document the INEX labs as part of CLEF 2012–2014. Table 1 gives an overview of the INEX activities over the year, including the specific tracks, with references to the respective overview papers.

#### 3.1 CLEF 2012

INEX 2012 (Bellot et al. 2012a) was the first INEX held as part of CLEF 2012, which was the eleventh annual edition of INEX. Recall that the last independent

**Table 1** CLEF INEX tracks over the years

| Tracks                  | 2012                     | 2013                   | 2014                  |
|-------------------------|--------------------------|------------------------|-----------------------|
| INEX overview           | Bellot et al. (2012a)    | Bellot et al. (2013a)  | Bellot et al. (2014a) |
| Linked data             | Wang et al. (2012)       | Gurajada et al. (2013) |                       |
| Relevance feedback      | Chappell and Geva (2012) |                        |                       |
| Snippet retrieval       | Trappett et al. (2012)   | Trappett et al. (2013) |                       |
| Social book search      | Koolen et al. (2012)     | Koolen et al. (2013)   | Koolen et al. (2014)  |
| Interactive SBS         |                          |                        | Hall et al. (2014)    |
| Tweet contextualization | SanJuan et al. (2012)    | Bellot et al. (2013b)  | Bellot et al. (2014b) |

INEX was held in December 2011, making the cycle to CLEF 2012 a particularly short year to run the full cycle of test collection development. INEX 2012 had again five tracks, all based on or derived from, the INEX 2011 tracks, which we will describe in this section.

*Linked Data (LD)* (Wang et al. 2012) was a new track but a direct descendant of the INEX 2011 Data-Centric Track, moving from the highly structured IMDb data to a rich textual corpus (Wikipedia) with rich semantic annotation (DBpedia).

**Linked Data Track** *investigating retrieval over a strongly structured collection of documents based on DBpedia and Wikipedia. The Ad Hoc Search Task has informational requests to be answered by the entities in DBpedia/Wikipedia. The Faceted Search Task asks for a restricted list of facets and facet-values that will optimally guide the searcher toward relevant information.*

LD did amazing efforts to create novel benchmarks at scale bringing linked data within the scope of IR. Perhaps this was too early, and the short nine-months cycle of INEX 2012 didn't help, as participation was small, and there were significant issues in creating the massive corpus merging Wikipedia with DBpedia and Yago linked data (but these issues were fixed in the next year).

*Relevance Feedback (RF)* (Chappell and Geva 2012) was a continuation of the track in INEX 2011, and the final edition of the track. The track studied (incremental) relevance feedback on the INEX Wikipedia Corpus from 2009 by reusing topics and judgments from earlier year. As it outgrew the Cranfield style evaluation effort, due to continuous submissions and scoring, it also lost some of the momentum of an annual cycle with clear deadlines, and it was decided not to continue it into 2013 but to offer it online to any interested party.

*Snippet Retrieval (SR)* (Trappett et al. 2012) was a continuing track of INEX 2011, investigating how to generate informative snippets for search results. This track also used the INEX Wikipedia Corpus from 2009, and reused some old ad hoc search topics plus created novel topics and judgments at both summary and document level. As the snippet retrieval track ran late in 2012—due to the short nine-month cycle of 2012—it was decided to carry it over to 2013, and build one final test collection for SR over the 2 years together.

*Social Book Search (SBS)* (Koolen et al. 2012) was another new track but a direct descendant of the INEX 2011 Books and Social Search Track, and the earlier Book Search Tracks (since 2007). Although the traditional out-of-copyright, full-text books still were continued as a task by special demand of those interested, the focus clearly shifted to the social book data, coming from Amazon and LibraryThing and originally constructed to support the INEX Interactive Track (running 2004–2010), including real-world complex book search requests, user profiles and personal book catalogues.

*Tweet Contextualization (TC)* (SanJuan et al. 2012) was also a new track, but directly derived from the INEX 2011 Question Answering Track, which focused on more NLP-oriented tasks and moved to multidocument summarization.

**Tweet Contextualization Track** *investigating tweet contextualization, answering questions of the form “what is this tweet about?” with a synthetic summary*

*of contextual information grasped from Wikipedia and evaluated by both the relevant text retrieved, and the “last point of interest.”*

The use case was based on short tweets or posts, which cannot be fully comprehensive, and to look for background information in Wikipedia about relevant entities, concepts, events, products, etc., referred to by the tweet or post. As the INEX Wikipedia corpus was getting dated, a new Wikipedia dump of 2012 was used. Topics were 1000 tweets, of which 63 were evaluated, both in terms of a synthetic measure against reference summaries, as well as by a human judgment on the readability of the whole summary.

At CLEF 2012 in Rome, the collocation with or embedding in the larger CLEF family worked out very well for INEX. There was a considerable new interest in the INEX tracks from other CLEF attendees, as INEX helped CLEF break free from the earlier focus on cross-language retrieval. During the CLEF organizers meeting there was active discussion on more direct collaboration between the different CLEF labs, which was strongly supported by the INEX organizers.

There are two major activities related to INEX that took place outside CLEF. First, there was the spin-off *Exploiting Semantic Annotations for Information Retrieval (ESAIR)* workshop that took place at CIKM 2012 at Maui (Kamps et al. 2013), featuring a range of papers and great keynotes on Knowledge Graphs (Evgeniy Gabrilovich), and Conversational Search (Ron Kaplan). This ESAIR workshop was continued from CIKM 2011 in Glasgow, and CIKM 2010 in Toronto. Second, a new track on *Contextual Suggestion* was started, running at the *Text REtrieval Conference (TREC)* 2012 (Dean-Hall et al. 2012), rather than at CLEF, to attract a wider attendance. This track was a direct result of the SIGIR 2011 workshop on *Supporting Complex Search Tasks: Entertain Me* (Belkin et al. 2011), which in turn was a spin-off of discussion at INEX 2010 (Beckers et al. 2010).

### 3.2 CLEF 2013

INEX 2013 (Bellot et al. 2013a) was the twelfth annual run of INEX, and the second as part of the CLEF family. With all the changes instigated in 2011 and 2012, INEX 2013 was a year of continuity, with training data for all tasks being available from 2012 and new, additional test collections being developed in 2013.

To better align with the general CLEF structure, INEX 2013 featured three *themes*: (1) searching professional and user generated data (SBS); searching structured or semantic data (LD); and focused retrieval (SR and TC). That is, INEX 2013 featured a total of four tasks, which we will discuss in the following.

*Linked Data (LD)* (Gurajada et al. 2013) managed to address the corpus issues of 2012, when a large part of the corpus was ultimately removed as it caused validity problems with the strict schema. This produced a massive linked data search corpus covering the core of the linked data graph (constituted by DBpedia) and rich textual sources with semantic annotations (the corresponding Wikipedia). The Faceted

Search task received less attention, due to the need for general queries with very large answer sets, rather than very selective queries expressing detailed information need. Of special mention is the new Jeopardy task, attempting to encourage writing or generating rich SPARQL style queries expressing complex needs. Although the LD track did not continue in 2014, the resulting corpus and test collections are a key resource and have been widely reused inside IR and beyond.

As explained in the section above, *Snippet Retrieval (SR)* (Trappett et al. 2013) was carried over from INEX 2012 as it ran so late for 2012 that it was decided to view this as early for 2013, effectively creating a 21 month cycle for INEX 2013. So the same topic set was used, additional runs were requested, and all judgments on snippet and document relevance were completed. As now a wealth of data for evaluating snippet retrieval was available, INEX 2013 presented the last year of the SR track.

*Social Book Search (SBS)* (Koolen et al. 2013) continued strong with the social book data. In particular the detailed statements of request, detailed user profiles and personal book catalogues, and expert book recommendations from the LibraryThing forums, proved very valuable data for research, as one of the few examples of social data being of general interest, and less fraught with privacy concerns. By popular demand of some participants, a small scale continuation of the “Prove It” task on the scanned book corpus was permitted by the track organizers, but the main focus was on the social book search tasks.

*Tweet Contextualization (TC)* (Bellot et al. 2013b) was picking up momentum, creating a test collection with 120 tweets over INEX 2012 and 2013. This was reflected by advance approaches, with the better systems combining NLP pipelines, twitter specific processing, and IR finesse, realizing the broader track’s goals to bridge IR and NLP approaches.

At CLEF 2013 in Valencia, there was a clear feeling of return-on-investments: all tracks were now stable and rerunning, and clearly attracting new participants from the broader CLEF community. At the same time, the increasing number of labs and tracks at CLEF led to an increasing competition for attention, and to thinning down the participation per track and task. This already led to the introduction of (or rather read: the restriction to) three themes in 2013 mentioned above, and continued in 2014 with pressure to reduce the number of tracks within INEX. This led the main organizers of INEX to discussions on whether to continue the INEX as a track, or split up the INEX tracks into 2–3 separate CLEF tracks, hence effectively discontinuing INEX. On a more positive note, intense discussion with the *Cultural Heritage in CLEF (CHiC)* lab led to the decision to fold CHiC into the Social Book Search Track of INEX in 2014, as an *Interactive Social Book Search (iSBS)* track.

There are three major activities related to INEX that took place outside CLEF. First, the spin-off ESAIR workshop continued at CIKM 2013 in San Francisco (Bennett et al. 2014), with a range of papers and keynotes on Wikification (Dan Roth), Reading difficulty annotation (Kevyn Collins-Thompson) and UI/UX for semantic annotations (Marti Hearst). Second, the spin-off *Contextual Suggestion Track* continued at TREC in 2013 (Dean-Hall et al. 2013). Third, a spin-off of the Social Book Search Track was the Structure Extraction (SE) task ran at ICDAR 2013

(Doucet et al. 2013), with the aim of evaluating automatic techniques for deriving structure from OCR and building hyperlinked table of contents.

### 3.3 CLEF 2014

INEX 2014 (Bellot et al. 2014a) was the thirteenth annual run of INEX, and the third as part of the CLEF family. Although INEX originally merged into CLEF as a “super lab” having a large number of tracks like ImageCLEF at the time, and now also absorbed the CLEF’s CHiC lab, it fully complied with the CLEF structure and limited itself to three tracks. Hence, INEX 2014 was more focused and featured three tasks, described now.

*Interactive Social Book Search (iSBS)* (Hall et al. 2014) was an exciting merge of the CHiC and INEX communities, with overwhelming support of the broader information science community, in order to run a large scale interactive track.

**Interactive Social Book Search Track** *investigating user information seeking behavior when interacting with various sources of information, for realistic task scenarios, and how the user interface impacts search and the search experience.*

The CHiC track (Petras et al. 2012, 2013) struggled with Europeana data, and with clear connections to the system-centric part of IR. INEX had run successful Interactive Tracks between 2004–2010 (Nordlie and Pharo 2012), but this line of activity lost momentum although the Amazon/LibraryThing corpus constructed for INEX iTrack experiments was still used in the SBS Track. This discussion led to federated effort of a large number of organizers to revive a user-centric interactive track at INEX, and to ensure it was seeded by the insights and challenges encountered in the system-centric SBS track. In particular, a novel type of multistage UI was developed that showed different functionality depending on the information seeking stage. Due to the relatively heavy system development in this inaugural year, a relatively modest number of 41 test persons participated in the track’s user studies.

*Social Book Search (SBS)* (Koolen et al. 2014) finally really ended the earlier scanned books tasks, and fully focused on the social book data also used in the iSBS track. The track boosted the number of user profiles and personal catalogues made available, far beyond those occurring in the search requests, in order to satisfy the taste of recommender system approaches. Of particular note is the seamless collaboration between the SBS and iSBS tracks, with extensive discussion between the two tracks, and the feeling that this discussion was interesting and helpful for both sides, realizing the dream of IR to fully integrate its computer science and information science parts.

*Tweet Contextualization (TC)* (Bellot et al. 2014b) continued strong, keeping the Wikipedia corpus stable since 2012 and 2013, but using a selection of tweets from the CLEF RepLab 2013, exploiting the extensive annotated data available for the RepLab tasks. The track was running in a similar way to 2013, except for the impact

of the changing character of the tweets used as topic, moving from general news to the specific product and services as used in the real-world social media monitoring part of RepLab. One of the main upshots of the track was that effective approaches for news tended to generalize to the specific realistic tweets derived from RepLab, demonstrating the clear value of the track's test collections. In addition to the main task in English, there was a pilot task in Spanish.

A fourth track, a continuation of the INEX Linked Data Track in collaboration with, and part of, the CLEF QA Track's QALD (QA from Linked Data) was announced, but the massive INEX corpus proved out of reach of the QALD participants and the desired collaboration didn't materialize in the end.

At CLEF 2014 in Sheffield, there were countless meetings between the iSBS and SBS track organizers, remindful of the endless discussions during the original INEX workshops at Dagstuhl. This led to an ambitious plan to boost interactive studies in a truly collaborative effort involving the largest group of organizers ever seen on any track, in any year, in any evaluation forum. . .

There are three major activities related to INEX that took place outside CLEF. First, the spin-off ESAIR workshop continued at CIKM 2014 in Shanghai (Alonso et al. 2015b), with a range of papers and keynotes on Semantic Search (Peter Mika) and Entity Linking (Silviu-Petru Cucerzan). Second, the spin-off *Contextual Suggestion Track* continued at TREC in 2014 (Dean-Hall et al. 2014). Third, a spin-off *New Trends in Content-Based Recommender Systems* (CBRecSys) workshop was held at RecSys 2014 (Bogers et al. 2014). As the SBS had content, ratings and catalogues but failed to attract significant numbers of researchers working on recommender systems to INEX and CLEF, a workshop was invented to feature the same data and tasks at RecSys. Although not many participated in the data challenge, the workshop was a resounding success and put content-based recommendation back on the agenda of RecSys as a first-class citizen.

### 3.4 End of INEX?

The INEX chairs decided not to submit a lab proposal for CLEF 2015. Instead, they strongly encouraged the remaining INEX tracks to submit their own proposals and become directly embedded into CLEF. First, thanks to the fruitful collaboration with CHiC and the former INEX *Interactive Track*, organizers, the INEX SBS track had great momentum and continued as a CLEF 2015 lab (to be further discussed in Sect. 4). Second, the INEX TC Track organizers were strongly advised to resubmit a significantly updated track proposal as a new CLEF lab, but this revision required more time, and it was decided to take a leap year for planning and reflection, and resubmit as a CLEF 2016 lab (to be further discussed in Sect. 5).

On a historical note, the INEX organizers already planned to step out and pass on the baton to the SBS track in 2014, as there was discussion (and some sense of agreement) to merge the TC track with CLEF RepLab. As in the end the RepLab/TC

merge didn't materialize, it was decided to continue with INEX for another year as the main CLEF lab label in 2014.

## 4 Social Book Search at CLEF

In this section, we document the *Social Book Search (SBS)* labs as part of CLEF 2015 and 2016. Table 2 gives an overview of the SBS activities over the year, including the specific tracks running, with references to the respective overview papers.

As detailed in the previous sections, the CLEF SBS lab stands on the shoulders of giants. It is a direct continuation of the SBS Track run at INEX since 2011, and has an even longer prehistory as the Book Search Track at INEX since 2007.

### 4.1 CLEF 2015

The CLEF 2015 *Social Book Search (SBS)* lab (Koolen et al. 2015a) was the first edition of the lab running as independent CLEF lab, and the fifth edition of SBS as part of the INEX family. As the CLEF SBS lab was a direct continuation of the INEX 2014 iSBS and SBS Tracks, there were two tracks running, which we will describe in further detail now.

The *Interactive Track* (Gäde et al. 2015a) was a direct continuation of the INEX iSBS track running as an extensive, concerted online user experiment.

**Interactive Track** *this is a user-centred track investigating how searchers use different types of metadata at various stages in the search process and how a search interface can support each stage in that process.*

An advanced system based on the Amazon/LibraryThing book collection was made available, with a baseline faceted search interface and multi-stage search interface exhibiting different functionalities depending on the information seeking stage.

The multistage interface worked with three different user interface configurations. In the start-up stage of search (pre-focus in the Kulthau/Vakkari model) where users explore the available information, the UI displays a *browse view*

**Table 2** CLEF social book search tracks over the years

| Tracks       | 2015                  | 2016                  |
|--------------|-----------------------|-----------------------|
| SBS overview | Koolen et al. (2015a) | Koolen et al. (2016a) |
| Suggestion   | Koolen et al. (2015b) | Koolen et al. (2016b) |
| Interactive  | Gäde et al. (2015a)   | Gäde et al. (2016)    |
| Mining       |                       | Bogers et al. (2016a) |



providing a query specific overview of the collection and navigation by Amazon subject classification as well as dense search results with title and ratings. In the main stage of search (focus in the Kulthau/Vakkari model) where users do an in-depth search and collect the relevant information, the UI displays the *search view* with a rich faceted search interface with more detail on each book result—this view is corresponding to advanced UIs as evolved in e-commerce and professional applications. In the final stage of search (post-focus in the Kulthau/Vakkari model) where users review and refine the selected information, and backtrack when needed, the UI displays the *book bag view* with all selected results and notes, plus providing a display of detailed information about each selected book.

In addition to the extensive, and innovative, system design and system building, about 200 test persons took part in the online user study comparing a tradition UI with the multi-stage UI, for both purposeful search and non-goal oriented search tasks, providing a very rich set of data for further analysis.

The *Suggestion Track* Koolen et al. (2015b) was the direct continuation of the INEX SBS track, focusing on IR for dealing with professional and user-generated data, and for exploring search that combines aspects of retrieval and content-based recommendation in a natural way.

**Suggestion Track** *this is a system-centred track focused on the comparative evaluation of systems in terms of how well they rank search results for complex book search requests that consist of both extensive natural language expressions of information needs as well as example books that reflect important aspects of those information needs, using a large collection of book descriptions with both professional metadata and user-generated content.*

The topics of 2015 were selected to include both a narrative statement of request (satisfying the needs of retrieval approaches) as well as one or more example books (satisfying the needs of recommender systems approaches). In addition, topics and book recommendations were humanly annotated, e.g., whether the example books or suggested books were positive (the requester wanted more like this, or responded positively recommended this), or negative, or neutral (no clear value judgement is expressed), in order to facilitate further analysis of the results. This led to an incredibly rich test collection, with a very high degree of realism as all requests and judgments are derived from LibraryThing forum data, supporting a wide range of experiments and deep analysis. Far more attention was given to appropriate recommender system evaluation, in order to attract those researchers to the track.

At CLEF 2015 in Toulouse, a large fraction of the many organizers were present, and there was a general sense of pride and content about the large efforts but also large achievements of this year. Of further special notice is the keynote on Polyrepresentation in Complex Book Search (Ingo Frommholz). The decision to spin off SBS as an independent CLEF lab clearly gave new impetus, and both tracks grew considerably.

There are four major activities related to SBS that took place outside CLEF.<sup>1</sup> First, the spin-off ESAIR workshop continued at CIKM 2015 in Melbourne (Balog et al. 2016). Second, the spin-off *Contextual Suggestion Track* continued at TREC 2015 (Dean-Hall et al. 2015). Third, the spin-off *New Trends in Content-Based Recommender Systems (CBRecSys)* workshop continued at RecSys 2015 (Bogers and Koolen 2015). Fourth, as mentioned above, a spin-off *Supporting Complex Search Tasks (SCST)* workshop was run at *European Conference on Information Retrieval (ECIR)* in Vienna (Gäde et al. 2015b). This workshop was a result of the CHiC and INEX interactive activities as part of the iSBS track, which wanted a mid-cycle deadline around ECIR, in order to finish the first round of activities timely, and reflect on the broader discussion and ways to take it forward.

## 4.2 CLEF 2016

The CLEF 2016 *Social Book Search (SBS)* lab (Koolen et al. 2016a) was the second run as independent CLEF lab, and sixth edition of SBS as part of the INEX family. The CLEF 2016 SBS lab continued the two existing tracks, and added a new track with a data mining/NLP focus. Hence, there were three tracks running, which we will describe in further detail now.

The *Interactive Track* Gäde et al. (2016) could cash in on all the development investments of the last 2 years, and ran a very similar track with minor refinements in the system and experimental setup, but with a large number of additional test persons in the user studies. One of the innovations was that also some of the LibraryThing requests from the forums (as used as topics in the Suggestion Track in the year before) were made part of the interactive experiments. Again a wealth of data was created (questionnaires and logs), but also again time to do a proper analysis of this rich data was running short, and regrettably only initial analysis was done and reported in the overview papers.

The *Mining Track* Bogers et al. (2016a) was a new addition to the track instigated by the observations in earlier years on the selection of suitable forum topics for use in the track, aiming to extract more information from the narrative part of the forums (or add other social media data).

**Mining Track** *this is a new track focused on detecting book search requests in forum posts for automatic book recommendation, as well as detecting and linking book titles in online book discussion forums.*

This led to two tasks, the first being the *book search request identification* task, in which the goal is to identify which threads on online forums are book search requests and locate the opening post with the actual request. This task directly

---

<sup>1</sup>In fact, there was also a related *Graph Search and Beyond (GSB)* workshop at SIGIR 2015 in Santiago de Chili (Alonso et al. 2015a), but this was a further spin-off of INEX rather than of SBS.

serviced a need of the track, as the process of selecting suitable topics, e.g., those containing a book request proper and enough book recommendations, was always done with a combination of scripts and manual inspection, that felt suboptimal. The data of earlier years presented suitable training data for effective classifiers. The second task was the *book linking* task, in which the goal is to recognize book titles in forum posts and link them to the corresponding metadata record through their unique book ID. This task was also in response to a direct need of the track, as it was observed that most, but not all, book suggestions can be extracted from the narrative part of the forum discussions, as not all book recommendations are properly annotated in the forum data. Again, the data of earlier years, hiding the explicitly annotated instances, provided ample training data for effective classifiers. These tasks would allow the track to go beyond the LibraryThing forums, and a large dump of Reddit book related discussion was part of the track, enabling the automatic classification of raw forum data into the format required by, or most useful for, the SBS track.

The *Suggestion Track* Koolen et al. (2016b) could also cash in on all the development investments of the last 2 years, and ran a very similar track with minor refinements in the setup, but with a large number of additional run and new groups joining the track. The innovations in the track setup were minor things making the setup even more perfect. To give an example, the removal of the few out of corpus recommendations: books recommended in the forum that are not in the Amazon collection, which doesn't affect the relative system ranking but gives slightly more accurate scores. Perhaps the main innovation was in the submitted systems, where the very rich data setup was used for the first time to train very specific word embeddings, that proved to be very effective.

The SBS sessions at CLEF 2016 in Évora featured strong attendance of participants and organizers, and included a keynote on the reception of literature and book search (Pertti Vakkari). In addition to the lab at CLEF, there are two major activities related to SBS that took place outside CLEF. First, the spin-off *Contextual Suggestion Track* continued at TREC 2016 (Hashemi et al. 2016). Second, the spin-off *New Trends in Content-Based Recommender Systems* (CBRecSys) workshop continued at RecSys 2016 (Bogers et al. 2016b).

### 4.3 To Be Continued?

At CLEF 2016 in Évora, there was considerable discussion on the future of the SBS lab, as no new lab or track proposal was submitted to CLEF. It was decided to take a sabbatical year for the track in 2017: with no track or lab running, but a year of reflection, with the intent to come back with a new lab proposal in 2018.

There was however, a second instance of the *Supporting Complex Search Tasks* (SCST) workshop, now held at the ACM SIGIR CHIIR Conference in Oslo (Koolen et al. 2017). The CHIIR workshop generate many new plans, and clearly

demonstrated the need for a track as iSBS which promotes direct collaboration between system-centric and user-centric researchers across the field of IR.

However, despite the many great plans, no concrete proposal was made for a 2018 track at CLEF, mostly due to staffing issues, as the same group of organizers was coordinating most of the work continuously for over a decade. The work on promoting and facilitating interactive IR systems and experiments does however continue with full force at the *Barriers to Interactive IR Resources Re-use (BIIRRR)* workshops held at the *Conference on Human Information Interaction and Retrieval (CHIIR)* in 2018 (Bogers et al. 2018) and 2019 (Bogers et al. 2019).

## 5 Cultural Microblog Contextualization at CLEF

In this section, we document the *Microblog Contextualization (MC2)* labs as part of CLEF 2016–2018. Table 3 gives an overview of the MC2 activities over the year, including the specific tracks running, and with references to the respective overview papers.

We include a brief discussion of the MC2 lab as it has a long prehistory as the INEX Tweet Contextualization track (2012–2014), which in turn was derived from the pre-CLEF INEX QA track since 2011.

### 5.1 CLEF 2016

The CLEF 2016 *Cultural Microblog Contextualization (MC2)* lab (Goeriot et al. 2016) was the first edition run as independent CLEF lab, and sixth edition as part of the INEX family.

After a year of reflection amongst the organizers of the INEX Tweet Contextualization Track in 2015, the track organizers submitted a significantly changed lab proposal to CLEF 2016, which was accepted as a workshop for CLEF 2016.

It is important to stress that *Cultural Microblog Contextualization (MC2)* lab was run as a workshop, and not as a regular CLEF track or lab. This despite its heavy focus on gathering, organizing, and delivering a relevant social data related to events

**Table 3** CLEF cultural microblog contextualization tracks over the years

| Tracks                              | 2016                   | 2017                    | 2018                 |
|-------------------------------------|------------------------|-------------------------|----------------------|
| MC2 overview                        | Goeriot et al. (2016)  | Ermakova et al. (2017a) | Hajjem et al. (2018) |
| Workshop report                     | Ermakova et al. (2016) |                         |                      |
| Content analysis                    |                        | Ermakova et al. (2017b) |                      |
| Search and timeline                 |                        | Goeriot et al. (2017)   |                      |
| Cross language/argumentative mining |                        |                         | Cossu et al. (2018)  |

generating a large number of micro-blog posts and web documents (such as, and in particular, cultural festivals). This resulted in an impressive amount of data and a pilot task, which are described in detail in (Ermakova et al. 2016).

This extensive corpus created in 2016 was used to support a proper MC2 lab in the following year.

## 5.2 CLEF 2017

The CLEF 2017 *Cultural Microblog Contextualization (MC2)* lab (Ermakova et al. 2017a) was the second edition run as independent CLEF lab, and seventh as part of the INEX family.

*Microblog Contextualization (MC2)* deals with how cultural context of a microblog affects its social impact at large. This involves microblog search, classification, filtering, language recognition, localization, entity extraction, linking open data and summarization. MC2 at CLEF 2017 featured three tasks.

*Content Analysis* (Ermakova et al. 2017b) was a new track on the NLP end of the spectrum, dealing with a number of classification tasks that are a prerequisite for other tasks involving noisy social media data.

**Content Analysis Track** *Given a stream of microblogs, filter out microblogs dealing with festivals and perform language(s) identification, event localization, author categorization, DBpedia entities recognition and automatic summarization of linked wikipedia pages in four languages.*

Some of these subtasks were based on the filtering and priority tasks of RepLab 2014.

*Microblog Search* (Goeuriot et al. 2017) was a new task to locate the most relevant microblog posts in the corpus, in response to a cultural query about festivals in Arabic, English, French, or Spanish.

**Microblog Search Track** *Given a cultural entity as a set of Wikipedia pages: i) retrieve relevant microblogs for an entity; or ii) summarize the most informative microblogs.*

The topics and queries were extracted in various ways from social media or review corpora.

*Timeline Illustration*, also documented in (Goeuriot et al. 2017), had the goal of retrieving all relevant tweets dedicated to each event of a festival, according to the program provided.

**Time Line Illustration Track** *The goal of the Timeline illustration based on Microblogs is to provide, for each event of a cultural festival, the most interesting tweets.*

The track focused on four large festivals, two music and two theater festivals in France and the UK.

### 5.3 CLEF 2018

The *Cultural Microblog Contextualization (MC2)* lab continues at CLEF 2018 (Hajjem et al. 2018), which is its third run as a CLEF lab, and eighth edition as part of the INEX family.

In 2018, the *Microblog Contextualization (MC2)* lab shifted its focus to *Multilingual Cultural Mining and Retrieval*. The lab promoted developing processing methods and resources to mine the social media sphere surrounding cultural events such as festivals. This requires to deal with almost all languages and dialects as well as informal expressions. A total of three tracks were organized in 2018.

The *cross language* task (Cossu et al. 2018) was specific to movies.

**Cross Language Cultural Retrieval over MicroBlogs Track** *investigating: (a) small microblogs multilingual information retrieval in Arabic, English, French and Latin languages; (b) microblogs bilingual information retrieval for tuning systems running on language pairs; (c) microblog monolingual information retrieval based on 2017 language identification.*

Topics were extracted from the French VodKaster website that allows readers to get personal short comments (or “microcritics”) about movies. The challenge of the task was to find related microblogs in four different languages in a large archive.

The *argumentation mining* task (Cossu et al. 2018) aimed to automatically identify reason-conclusion structures from text, which can model the position, stance or attitude (as expressed via Twitter microblogs) of a social web user about a cultural event.

**Mining Opinion Argumentation Track** *investigating: (a) polarity detection in microblogs; (b) automatic identification of argumentation elements over microblogs and Wikipedia; (c) classification and summarization of arguments in texts.*

The idea was to perform a search process on a massive microblog collection that focuses on claims about a given festival.

In addition, there was a new pilot task on dialect or language variation detection using a new corpus, and extending the earlier 2017 language recognition task.

**Dialectal Focus Retrieval Track** *investigating: (a) Arabic dialects in blogs, microblogs and video news transcriptions; (b) Spanish language variations in blogs, microblog and journals.*

### 5.4 To Be Continued?

At CLEF 2018 in Avignon, there was extensive discussion on the future of the *Cultural Microblog Contextualization (MC2)* lab as no new lab or track proposal was submitted to CLEF 2019. The track addressed an important area of research, and

managed to attract a strong base of participants with ongoing research and interest in this area, and plans for follow-up activities within CLEF or elsewhere are ongoing. The main result of the *Microblog Contextualization (MC2)* tracks, however, is a great number of unique test collections that can be used for future experiments.

## 6 Discussion and Conclusions

This completes our rundown of the *INitiative for the Evaluation of XML Retrieval (INEX)* lab of CLEF 2012–2014, and follow up CLEF *Social Book Search (SBS)* lab (2015–2016) and CLEF *Cultural Microblog Contextualization (MC2)* lab (2016–2018). We emphasize the INEX and SBS activities, as it's too early to look back on the MC2 lab which is still in full swing at the time of writing.

As we promised in the introduction, this chapter doesn't provide a definite account of these labs, as no single chapter can do justice to the massive amount of work done in the INEX and follow up labs. Rather, we hope this chapter to be useful as a starting point with references to the respective overview papers providing full details on the labs, tracks and resulting test collections. We hope that the comprehensive high level overview helps convey the impressive breadth and scope of activities spanning many years, trying to highlight some relations between various activities inside and outside CLEF that may not be immediately apparent, thereby also highlighting some of the important developments in the field during these years.

There is one aspect in which this chapter is significantly lacking: one of the main impacts of the labs is in all the research papers it enabled and encouraged: the numerous track participation papers in the CLEF working notes, the many conference papers derived from track participation, both in the CLEF conference proceedings as in other proceedings and journals, as well as all the other papers that use some of benchmarks. As a case in point, Google Scholar lists over well over four thousand papers that mention INEX. Understanding the impact of INEX in particular, and CLEF in general, would need to take this research uptake into account.

One of INEX's key contributions to IR is that it has been completely volunteer run since 2002, with all organization and activities crowdsourced to the participants: from the start (proposal of tracks and tasks) until the end (topic creation, topic assessment). This created a generation of researchers that were touched by INEX, and all contributed to its great success. Just to mention a selection of key organizers involved in 2012–2018: *P. Bellot, T. Bogers, T. Chappell, J. Cossu, A. Doucet, L. Ermakova, S. Geva, L. Goeuriot, S. Gurajada, M. Gäde, M. Hajjem, M. A. Hall, I. Hendrickx, H. C. Huurdeman, J. Kamps, G. Kazai, M. Koolen, M. Landoni, M. Marx, A. Mishra, V. Moriceau, J. Mothe, P. Mulhem, J.-N. Nie, M. Preminger, G. Ramírez, E. SanJuan, M. Sanderson, E. Sanjuan, R. Schenkel, F. Scholer, A. Schuh, M. Skov, X. Tannier, M. Theobald, E. Toms, M. Trappett, A. Trotman, S. Verberne, D. Walsh, and Q. Wang.* But there were countless more participant volunteers

helping out with the tracks, and becoming part of the extended INEX family. It is through this generation of researchers, and their follow-up students, that INEX is making a lasting impact on the field for many years to come.

**Acknowledgements** Thanks to the CLEF Association, CLEF Conference, and CLEF Labs organizers for the wonderful support over the years, that greatly facilitated all the work reported in this chapter. Special thanks to the editors of this volume for support, encouragement and great flexibility.

## References

- Alonso O, Hearst MA, Kamps J (2015a) Report on the first SIGIR workshop on graph search and beyond (GSB'15). SIGIR Forum 49(2):89–97. <http://doi.acm.org/10.1145/2888422.2888436>
- Alonso O, Kamps J, Karlgren J (2015b) Report on the seventh workshop on exploiting semantic annotations in information retrieval (ESAIR'14). SIGIR Forum 49(1):27–34. <http://doi.acm.org/10.1145/2795403.2795412>
- Balog K, Dalton J, Doucet A, Ibrahim Y (2016) Report on the eighth workshop on exploiting semantic annotations in information retrieval (ESAIR '15). SIGIR Forum 50(1):49–57. <http://doi.acm.org/10.1145/2964797.2964806>
- Beckers T, Bellot P, Demartini G, Denoyer L, Vries CMD, Doucet A, Fachry KN, Fuhr N, Gallinari P, Geva S, Huang WC, Iofciu T, Kamps J, Kazai G, Koolen M, Kutty S, Landoni M, Lehtonen M, Moriceau V, Nayak R, Nordlie R, Pharo N, SanJuan E, Schenkel R, Tannier X, Theobald M, Thom JA, Trotman A, de Vries AP (2010) Report on INEX 2009. SIGIR Forum 44(1):38–57. <http://doi.acm.org/10.1145/1842890.1842897>
- Belkin NJ, Clarke CLA, Gao N, Kamps J, Karlgren J (2011) Report on the SIGIR workshop on “entertain me”: supporting complex search tasks. SIGIR Forum 45(2):51–59. <http://doi.acm.org/10.1145/2093346.2093354>
- Bellot P, Chappell T, Doucet A, Geva S, Gurajada S, Kamps J, Kazai G, Koolen M, Landoni M, Marx M, Mishra A, Moriceau V, Mothe J, Preminger M, Ramírez G, Sanderson M, Sanjuan E, Scholer F, Schuh A, Tannier X, Theobald M, Trappett M, Trotman A, Wang Q (2012a) Report on INEX 2012. SIGIR Forum 46(2):50–59. <http://doi.acm.org/10.1145/2422256.2422264>
- Bellot P, Chappell T, Doucet A, Geva S, Kamps J, Kazai G, Koolen M, Landoni M, Marx M, Moriceau V, Mothe J, Ramírez G, Sanderson M, SanJuan E, Scholer F, Tannier X, Theobald M, Trappett M, Trotman A, Wang Q (2012b) Report on INEX 2011. SIGIR Forum 46(1):33–42. <http://doi.acm.org/10.1145/2215676.2215679>
- Bellot P, Doucet A, Geva S, Gurajada S, Kamps J, Kazai G, Koolen M, Mishra A, Moriceau V, Mothe J, Preminger M, SanJuan E, Schenkel R, Tannier X, Theobald M, Trappett M, Wang Q (2013a) Overview of INEX 2013. In: Forner P, Müller H, Paredes R, Rosso P, Stein B (eds) Information access evaluation meets multilinguality, multimodality, and visualization. Proceedings of the fourth international conference of the CLEF initiative (CLEF 2013). Lecture Notes in Computer Science (LNCS) 8138. Springer, Heidelberg, pp 269–281
- Bellot P, Moriceau V, Mothe J, SanJuan E, Tannier X (2013b) Overview of INEX tweet contextualization 2013 track. In: Forner P, Navigli R, Tufis D, Ferro N (eds) Working notes for CLEF 2013 conference, Valencia, Spain, September 23–26, 2013. CEUR-WS.org. CEUR workshop proceedings, vol 1179. <http://ceur-ws.org/Vol-1179/CLEF2013wn-INEX-BellotEt2013.pdf>



- Bellot P, Bogers T, Geva S, Hall MA, Huurdeman HC, Kamps J, Kazai G, Koolen M, Moriceau V, Mothe J, Preminger M, SanJuan E, Schenkel R, Skov M, Tannier X, Walsh D (2014a) Overview of INEX 2014. In: Kanoulas E, Lupu M, Clough P, Sanderson M, Hall M, Hanbury A, Toms E (eds) Information access evaluation – multilinguality, multimodality, and interaction. Proceedings of the fifth international conference of the CLEF initiative (CLEF 2014). Lecture Notes in Computer Science (LNCS) 8685. Springer, Heidelberg, pp 212–228
- Bellot P, Moriceau V, Mothe J, SanJuan E, Tannier X (2014b) Overview of INEX tweet contextualization 2014 track. In: Cappellato L, Ferro N, Halvey M, Kraaij W (eds) Working notes for CLEF 2014 conference, Sheffield, UK, September 15–18, 2014. CEUR-WS.org. CEUR Workshop proceedings, vol 1180, pp 494–500. <http://ceur-ws.org/Vol-1180/CLEF2014wn-BellotEt2014.pdf>
- Bennett PN, Gabrilovich E, Kamps J, Karlgren J (2014) Report on the sixth workshop on exploiting semantic annotations in information retrieval (ESAIR'13). SIGIR Forum 48(1):13–20. <http://doi.acm.org/10.1145/2641383.2641387>
- Bogers T, Koolen M (2015) Second workshop on new trends in content-based recommender systems (cbrecsys 2015). In: Werthner H, Zanker M, Golbeck J, Semeraro G (eds) Proceedings of the 9th ACM conference on recommender systems, RecSys 2015, Vienna, Austria, September 16–20, 2015. ACM, New York, pp 339–340. <http://doi.acm.org/10.1145/2792838.2798718>
- Bogers T, Koolen M, Cantador I (2014) Workshop on new trends in content-based recommender systems: (cbrecsys 2014). In: Kobsa A, Zhou MX, Ester M, Koren Y (eds) Eighth ACM conference on recommender systems, RecSys '14, Foster City, Silicon Valley, CA, USA - October 06–10, 2014. ACM, New York, pp 379–380. <http://doi.acm.org/10.1145/2645710.2645784>
- Bogers T, Hendrickx I, Koolen M, Verberne S (2016a) Overview of the SBS 2016 mining track. In: Balog K, Cappellato L, Ferro N, Macdonald C (eds) Working notes of CLEF 2016 - conference and labs of the evaluation forum, Évora, Portugal, 5–8 September, 2016. CEUR-WS.org, CEUR workshop proceedings, vol 1609, pp 1053–1063. <http://ceur-ws.org/Vol-1609/16091053.pdf>
- Bogers T, Koolen M, Musto C, Lops P, Semeraro G (2016b) Third workshop on new trends in content-based recommender systems (cbrecsys 2016). In: Sen S, Geyer W, Freyne J, Castells P (eds) Proceedings of the 10th ACM conference on recommender systems, Boston, MA, USA, September 15–19, 2016. ACM, New York, pp 419–420. <http://doi.acm.org/10.1145/2959100.2959200>
- Bogers T, Gäde M, Hall MM, Freund L, Koolen M, Petras V, Skov M (2018) Report on the workshop on barriers to interactive IR resources re-use (BIIRRR 2018). SIGIR Forum 52(1):119–128. <https://doi.org/10.1145/3274784.3274795>
- Bogers T, Dodson S, Freund L, Gäde M, Hall MM, Koolen M, Petras V, Pharo N, Skov M (2019) Workshop on barriers to interactive IR resources re-use (BIIRRR 2019). In: Azzopardi L, Halvey M, Ruthven I, Joho H, Murdock V, Qvarfordt P (eds) Proceedings of the 2019 conference on human information interaction and retrieval, CHIIR 2019, Glasgow, Scotland, UK, March 10–14, 2019. ACM, New York, pp 389–392. <https://doi.org/10.1145/3295750.3298965>
- Chappell T, Geva S (2012) Overview of the INEX 2012 relevance feedback track. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) CLEF 2012 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Cossu J, Gonzalo J, Hajjem M, Hamon O, Latiri C, SanJuan E (2018) CLEF MC2 2018 lab technical overview of cross language microblog search and argumentative mining. In: Cappellato L, Ferro N, Nie J, Soulier L (eds) Working notes of CLEF 2018 - conference and labs of the evaluation forum, Avignon, France, September 10–14, 2018. CEUR-WS.org. CEUR workshop proceedings, vol 2125, [http://ceur-ws.org/Vol-2125/invited\\_paper\\_7.pdf](http://ceur-ws.org/Vol-2125/invited_paper_7.pdf)
- Dean-Hall A, Clarke CLA, Kamps J, Thomas P, Voorhees EM (2012) Overview of the TREC 2012 contextual suggestion track. In: Voorhees EM, Buckland LP (eds) Proceedings of the twenty-first text retrieval conference, TREC 2012, Gaithersburg, Maryland, USA, November 6–9, 2012. National Institute of Standards and Technology (NIST), Gaithersburg, vol Special Publication 500-298, <http://trec.nist.gov/pubs/trec21/papers/CONTEXTUAL12.overview.pdf>

- Dean-Hall A, Clarke CLA, Simone N, Kamps J, Thomas P, Voorhees EM (2013) Overview of the TREC 2013 contextual suggestion track. In: Voorhees EM (ed) Proceedings of the twenty-second text retrieval conference, TREC 2013, Gaithersburg, Maryland, USA, November 19–22, 2013. National Institute of Standards and Technology (NIST), Gaithersburg, vol Special Publication 500-302 <http://trec.nist.gov/pubs/trec22/papers/CONTEXT.OVERVIEW.pdf>
- Dean-Hall A, Clarke CLA, Kamps J, Thomas P, Voorhees EM (2014) Overview of the TREC 2014 contextual suggestion track. In: Voorhees EM, Ellis A (eds) Proceedings of the twenty-third text retrieval conference, TREC 2014, Gaithersburg, Maryland, USA, November 19–21, 2014. National Institute of Standards and Technology (NIST), Gaithersburg, vol Special Publication 500-308, <http://trec.nist.gov/pubs/trec23/papers/overview-context.pdf>
- Dean-Hall A, Clarke CLA, Kamps J, Kiseleva J, Voorhees EM (2015) Overview of the TREC 2015 contextual suggestion track. In: Voorhees EM, Ellis A (eds) Proceedings of the twenty-fourth text retrieval conference, TREC 2015, Gaithersburg, Maryland, USA, November 17–20, 2015. National Institute of Standards and Technology (NIST), Gaithersburg, vol Special Publication 500-319 <http://trec.nist.gov/pubs/trec24/papers/Overview-CX.pdf>
- Doucet A, Kazai G, Colutto S, Mühlberger G (2013) ICDAR 2013 competition on book structure extraction. In: 12th international conference on document analysis and recognition, ICDAR 2013, Washington, DC, USA, August 25–28, 2013, IEEE Computer Society, Washington, pp 1438–1443. <https://doi.org/10.1109/ICDAR.2013.290>
- Ermakova L, Goeriot L, Mothe J, Mulhem P, Nie J, SanJuan E (2016) Cultural micro-blog contextualization 2016 workshop overview: data and pilot tasks. In: Balog K, Cappellato L, Ferro N, Macdonald C (eds) Working notes of CLEF 2016 - conference and labs of the evaluation forum, Évora, Portugal, 5–8 September, 2016. CEUR-WS.org. CEUR workshop proceedings, vol 1609, pp 1197–1200. <http://ceur-ws.org/Vol-1609/16091197.pdf>
- Ermakova L, Goeriot L, Mothe J, Mulhem P, Nie J, SanJuan E (2017a) CLEF 2017 microblog cultural contextualization lab overview. In: Jones GJF, Lawless S, Gonzalo J, Kelly L, Goeriot L, Mandl T, Cappellato L, Ferro N (eds) Experimental IR meets multilinguality, multimodality, and interaction - 8th international conference of the CLEF association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings. Lecture Notes in Computer Science, vol 10456, pp 304–314. Springer, Basel. [https://doi.org/10.1007/978-3-319-65813-1\\_27](https://doi.org/10.1007/978-3-319-65813-1_27)
- Ermakova L, Mothe J, SanJuan E (2017b) CLEF 2017 microblog cultural contextualization content analysis task overview. In: Cappellato L, Ferro N, Goeriot L, Mandl T (eds) Working notes of CLEF 2017 - conference and labs of the evaluation forum, Dublin, Ireland, September 11–14, 2017. CEUR-WS.org, CEUR workshop proceedings, vol 1866. [http://ceur-ws.org/Vol-1866/invited\\_paper\\_14.pdf](http://ceur-ws.org/Vol-1866/invited_paper_14.pdf)
- Gäde M, Hall MM, Huurdeman HC, Kamps J, Koolen M, Skov M, Toms E, Walsh D (2015a) Overview of the SBS 2015 interactive track. In: Cappellato L, Ferro N, Jones GJF, SanJuan E (eds) Working notes of CLEF 2015 - conference and labs of the evaluation forum, Toulouse, France, September 8–11, 2015. CEUR-WS.org, CEUR workshop proceedings, vol 1391. <http://ceur-ws.org/Vol-1391/78-CR.pdf>
- Gäde M, Hall MM, Huurdeman HC, Kamps J, Koolen M, Skov M, Toms E, Walsh D (2015b) Report on the first workshop on supporting complex search tasks. SIGIR Forum 49(1):50–56. <http://doi.acm.org/10.1145/2795403.2795415>
- Gäde M, Hall MM, Huurdeman HC, Kamps J, Koolen M, Skov M, Bogers T, Walsh D (2016) Overview of the SBS 2016 interactive track. In: Balog K, Cappellato L, Ferro N, Macdonald C (eds) Working notes of CLEF 2016 - conference and labs of the evaluation forum, Évora, Portugal, 5–8 September, 2016. CEUR-WS.org. CEUR workshop proceedings, vol 1609, pp 1024–1038. <http://ceur-ws.org/Vol-1609/16091024.pdf>
- Goeriot L, Mothe J, Mulhem P, Murtagh F, SanJuan E (2016) Overview of the CLEF 2016 cultural micro-blog contextualization workshop. In: Fuhr N, Quaresma P, Gonçalves T, Larsen B, Balog K, Macdonald C, Cappellato L, Ferro N (eds) Experimental IR meets multilinguality, multimodality, and interaction. Proceedings of the seventh international conference of the CLEF association (CLEF 2016), Lecture notes in computer science (LNCS) 9822. Springer, Heidelberg, pp 371–378

- Goeriot L, Mulhem P, SanJuan E (2017) CLEF 2017 MC2 search and time line tasks overview. In: Cappellato L, Ferro N, Goeriot L, Mandl T (eds) Working notes of CLEF 2017 - conference and labs of the evaluation forum, Dublin, Ireland, September 11–14, 2017. CEUR-WS.org, CEUR workshop proceedings, vol 1866. [http://ceur-ws.org/Vol-1866/invited\\_paper\\_15.pdf](http://ceur-ws.org/Vol-1866/invited_paper_15.pdf)
- Gurajada S, Kamps J, Mishra A, Schenkel R, Theobald M, Wang Q (2013) Overview of the INEX 2013 linked data track. In: Forner P, Navigli R, Tufis D, Ferro N (eds) Working notes for CLEF 2013 conference, Valencia, Spain, September 23–26, 2013. CEUR-WS.org, CEUR workshop proceedings, vol 1179. <http://ceur-ws.org/Vol-1179/CLEF2013wn-INEX-GurajadaEt2013.pdf>
- Hajjem M, Cossu J, Latiri C, SanJuan E (2018) CLEF MC2 2018 lab overview. In: Bellot P, Trabelsi C, Mothe J, Murtagh F, Nie J, Soulier L, SanJuan E, Cappellato L, Ferro N (eds) Experimental IR meets multilinguality, multimodality, and interaction - 9th international conference of the CLEF association, CLEF 2018, Avignon, France, September 10–14, 2018, Proceedings. Lecture Notes in Computer Science, vol 11018, pp 302–308. Springer. [https://doi.org/10.1007/978-3-319-98932-7\\_27](https://doi.org/10.1007/978-3-319-98932-7_27)
- Hall MM, Huurdeman HC, Koolen M, Skov M, Walsh D (2014) Overview of the INEX 2014 interactive social book search track. In: Cappellato L, Ferro N, Halvey M, Kraaij W (eds) Working notes for CLEF 2014 conference, Sheffield, UK, September 15–18, 2014. CEUR-WS.org, CEUR workshop proceedings, vol 1180, pp 480–493. <http://ceur-ws.org/Vol-1180/CLEF2014wn-Inex-HallEt2014.pdf>
- Hashemi SH, Clarke CLA, Kamps J, Kiseleva J, Voorhees EM (2016) Overview of the TREC 2016 contextual suggestion track. In: Voorhees EM, Ellis A (eds) Proceedings of the twenty-fifth text retrieval conference, TREC 2016, Gaithersburg, Maryland, USA, November 15–18, 2016, National Institute of Standards and Technology (NIST), Gaithersburg, vol Special Publication 500-321. <http://trec.nist.gov/pubs/trec25/papers/Uamsterdam-CX.pdf>
- Kamps J, Geva S, Peters C, Sakai T, Trotman A, Voorhees EM (2009) Report on the SIGIR 2009 workshop on the future of IR evaluation. SIGIR Forum 43(2):13–23. <http://doi.acm.org/10.1145/1670564.1670567>
- Kamps J, Karlgren J, Mika P, Murdock V (2013) Report on the fifth workshop on exploiting semantic annotations in information retrieval (ESAIR'12). SIGIR Forum 47(1):38–45. <http://doi.acm.org/10.1145/2492189.2492196>
- Kazai G (2009, 2018) INEX. In: Liu L, Özsu MT (eds) Encyclopedia of database systems. Springer, New York, p 1472. [https://doi.org/10.1007/978-0-387-39940-9\\_2846](https://doi.org/10.1007/978-0-387-39940-9_2846) [https://doi.org/10.1007/978-0-387-39940-9\\_2846](https://doi.org/10.1007/978-0-387-39940-9_2846)
- Koolen M, Kazai G, Kamps J, Preminger M, Doucet A, Landoni M (2012) Overview of the INEX 2012 social book search track. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) CLEF 2012 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Koolen M, Kazai G, Preminger M, Doucet A (2013) Overview of the INEX 2013 social book search track. In: Forner P, Navigli R, Tufis D, Ferro N (eds) Working notes for CLEF 2013 conference, Valencia, Spain, September 23–26, 2013. CEUR-WS.org, CEUR workshop proceedings, vol 1179. <http://ceur-ws.org/Vol-1179/CLEF2013wn-INEX-KoolenEt2013b.pdf>
- Koolen M, Bogers T, Kamps J, Kazai G, Preminger M (2014) Overview of the INEX 2014 social book search track. In: Cappellato L, Ferro N, Halvey M, Kraaij W (eds) Working notes for CLEF 2014 conference, Sheffield, UK, September 15–18, 2014. CEUR-WS.org, CEUR workshop proceedings, vol 1180, pp 462–479. <http://ceur-ws.org/Vol-1180/CLEF2014wn-Inex-KoolenEt2014.pdf>
- Koolen M, Bogers T, Gäde M, Hall MA, Huurdeman HC, Kamps J, Skov M, Toms E, Walsh D (2015a) Overview of the CLEF 2015 social book search lab. In: Mothe J, Savoy J, Kamps J, Pinel-Sauvagnat K, Jones GJF, SanJuan E, Cappellato L, Ferro N (eds) Experimental IR meets multilinguality, multimodality, and interaction. Proceedings of the sixth international conference of the CLEF association (CLEF 2015). Lecture notes in computer science (LNCS) 9283. Springer, Heidelberg, pp 545–564

- Koolen M, Bogers T, Kamps J (2015b) Overview of the SBS 2015 suggestion track. In: Cappellato L, Ferro N, Jones GJF, SanJuan E (eds) Working notes of CLEF 2015 - conference and labs of the evaluation forum, Toulouse, France, September 8–11, 2015. CEUR-WS.org. CEUR workshop proceedings, vol 1391. <http://ceur-ws.org/Vol-1391/76-CR.pdf>
- Koolen M, Bogers T, Gäde M, Hall M, Hendrickx I, Huurdeman HC, Kamps J, Skov M, Verberne S, Walsh D (2016a) Overview of the CLEF 2016 social book search lab. In: Fuhr N, Quaresma P, Gonçalves T, Larsen B, Balog K, Macdonald C, Cappellato L, Ferro N (eds) Experimental IR meets multilinguality, multimodality, and interaction. Proceedings of the seventh international conference of the CLEF association (CLEF 2016). Lecture notes in computer science (LNCS) 9822. Springer, Heidelberg, pp 351–370
- Koolen M, Bogers T, Kamps J (2016b) Overview of the SBS 2016 suggestion track. In: Balog K, Cappellato L, Ferro N, Macdonald C (eds) Working notes of CLEF 2016 - conference and labs of the evaluation forum, Évora, Portugal, 5–8 September, 2016. CEUR-WS.org, CEUR workshop proceedings, vol 1609, pp 1039–1052. <http://ceur-ws.org/Vol-1609/16091039.pdf>
- Koolen M, Kamps J, Bogers T, Belkin NJ, Kelly D, Yilmaz E (2017) Report on the second workshop on supporting complex search tasks. SIGIR Forum 51(1):58–66. <http://doi.acm.org/10.1145/3130332.3130343>
- Nordlie R, Pharo N (2012) Seven years of INEX interactive retrieval experiments - lessons and challenges. In: Catarci T, Forner P, Hiemstra D, Peñas A, Santucci G (eds) Information access evaluation. multilinguality, multimodality, and visual analytics. Proceedings of the third international conference of the CLEF initiative (CLEF 2012). Lecture notes in computer science (LNCS) 7488. Springer, Heidelberg, pp 13–23
- Petras V, Ferro N, Gäde M, Isaac A, Kleineberg M, Masiero I, Nicchio M, Stiller J (2012) Cultural heritage in CLEF (CHiC) overview 2012. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) CLEF 2012 working notes, CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Petras V, Bogers T, Hall M, Savoy J, Malak P, Pawlowski A, Ferro N, Masiero I (2013) Cultural heritage in CLEF (CHiC) 2013. In: Forner P, Müller H, Paredes R, Rosso P, Stein B (eds) Information access evaluation meets multilinguality, multimodality, and visualization. Proceedings of the fourth international conference of the CLEF initiative (CLEF 2013). Lecture notes in computer science (LNCS) 8138. Springer, Heidelberg, pp 192–211
- SanJuan E, Moriceau V, Tannier X, Bellot P, Mothe J (2012) Overview of the INEX 2012 tweet contextualization track. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) CLEF 2012 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Thanos C, Casarosa V (2017) The key role of the DELOS network of excellence in establishing digital libraries as a research field in Europe. LIBER Quart 26(4):296–307. <http://doi.org/10.18352/lq.10165>
- Trappett M, Geva S, Trotman A, Scholer F, Sanderson M (2012) Overview of the INEX 2012 snippet retrieval track. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) CLEF 2012 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>
- Trappett M, Geva S, Trotman A, Scholer F, Sanderson M (2013) Overview of the INEX 2013 snippet retrieval track. In: Forner P, Navigli R, Tufis D, Ferro N (eds) Working notes for CLEF 2013 conference, Valencia, Spain, September 23–26, 2013. CEUR-WS.org, CEUR workshop proceedings, vol 1179. <http://ceur-ws.org/Vol-1179/CLEF2013wn-INEX-TrappettEt2013.pdf>
- Wang Q, Kamps J, Ramírez Camps G, Marx M, Schuth A, Theobald M, Gurajada S, Mishra A (2012) Overview of the INEX 2012 linked data track. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) CLEF 2012 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1178/>

**Part V**  
**Beyond Retrieval**

# Results and Lessons of the Question Answering Track at CLEF



**Anselmo Peñas, Álvaro Rodrigo, Bernardo Magnini, Pamela Forner, Eduard Hovy, Richard Sutcliffe, and Danilo Giampiccolo**

**Abstract** The Question Answering track at CLEF ran for 13 years, from 2003 until 2015. Along these years, many different tasks, resources and evaluation methodologies were developed. We divide the CLEF Question Answering campaigns into four eras: (1) Ungrouped mainly factoid questions asked against monolingual newspapers (2003–2006), (2) Grouped questions asked against newspapers and Wikipedias (2007–2008), (3) Ungrouped questions against multilingual parallel-aligned EU legislative documents (2009–2010), and (4) Questions about a single document using a related document collection as background information (2011–2015). We provide the description and the main results for each of these eras, together with the pilot exercises and other Question Answering tasks that ran in CLEF. Finally, we conclude with some of the lessons learnt along these years.

---

A. Peñas · Á. Rodrigo (✉)

NLP&IR Group at UNED, Madrid, Spain  
e-mail: [anselmo@lsi.uned.es](mailto:anselmo@lsi.uned.es); [alvarory@lsi.uned.es](mailto:alvarory@lsi.uned.es)

B. Magnini

Natural Language Processing Research Unit, FBK, Trento, Italy  
e-mail: [magnini@fbk.eu](mailto:magnini@fbk.eu)

P. Forner · D. Giampiccolo

FBK - PMG, Trento, Italy  
e-mail: [forner@fbk.eu](mailto:forner@fbk.eu); [giampiccolo@fbk.eu](mailto:giampiccolo@fbk.eu)

E. Hovy

Carnegie Mellon University, Language Technologies Institute, Pittsburgh, PA, USA  
e-mail: [hovy@cmu.edu](mailto:hovy@cmu.edu)

R. Sutcliffe

CSIS Department, University of Limerick, Limerick, Ireland  
e-mail: [Richard.Sutcliffe@ul.ie](mailto:Richard.Sutcliffe@ul.ie)

© Springer Nature Switzerland AG 2019

N. Ferro, C. Peters (eds.), *Information Retrieval Evaluation in a Changing World*, The Information Retrieval Series 41,  
[https://doi.org/10.1007/978-3-030-22948-1\\_18](https://doi.org/10.1007/978-3-030-22948-1_18)

## 1 Introduction

Under the promotion of the TREC-8 (Voorhees and Tice 1999) and TREC-9 (Voorhees 2000) Question Answering tracks, research in Question Answering (QA) received a strong boost. The aim of the TREC QA campaigns was to assess the capability of systems to return exact answers to open-domain English questions. The QA track at TREC represented the first attempt to foster and emphasise the importance of research on systems that could extract relevant and precise information from textual documents rather than retrieve and rank these documents. QA systems were designed to find answers to open domain questions in a large collection of documents and the development of such systems has acquired an important status among the scientific community because it entails research in both Natural Language Processing (NLP) and Information Retrieval (IR), putting the two disciplines in contact. In contrast to the IR scenario, a QA system processes questions formulated into natural language (instead of keyword-based queries) and retrieves answers (instead of documents). During the years at TREC from 1999 to 2007 and under the TAC conference in 2008, the task evolved, providing advancements and evaluation evidence for a number of key aspects in QA, including answering factual and definition questions, questions requiring complex analysis, follow-up questions in a dialog-like context, and mining answers from different text genres, including blogs.

In this context, research and evaluation of Question Answering systems were promoted in Europe by CLEF for two reasons: (1) to deal with QA in other languages besides English and (2) to deal with multilingual scenarios. Multilingual QA emerged as a complementary research task, representing a promising direction for at least two reasons. First, it allowed users to interact with machines in their native languages, contributing to easier, faster, and more equal information access. Second, cross-lingual capabilities enabled QA systems to access information stored only in language-specific text collections.

In a multilingual QA task two main variables need to be considered: (1) the source language, i.e. the language in which the questions are formulated, and (2) the target language, i.e. the language of the document collection. A cross-language QA system should enable users to search documents that are written in a language they do not know, which is a promising application in a multilingual society. Answer strings, which are usually retrieved from the corpus without any changes, could be translated into the source language, but this further cross-lingual step was not required in the track.

During the years, the effort of the QA track at CLEF organisers was focused on two main issues. One aim was to offer an evaluation exercise characterised by cross-linguality, covering as many languages as possible. From this perspective, major attention was given to European languages, adding at least one new language each year. However, the offer was also kept open to languages from all over the world, as the inclusion of Indonesian shows.

The other important issue was to maintain a balance between the established procedure inherited from the TREC campaigns and innovation. This allowed newcomers to join the competition and, at the same time, offered “veterans” new challenges.

The QA campaigns can be divided into four eras:

- Era I: 2003–2006. Ungrouped mainly factoid questions asked against monolingual newspapers; Exact answers returned.
- Era II: 2007–2008. Grouped questions asked against newspapers and Wikipedias; Exact answers returned.
- Era III: 2009–2010. Ungrouped questions against multilingual parallel-aligned EU legislative documents; Passages or exact answers returned.
- Era IV: 2011–2015. Questions about a single document using a related document collection as background information. Multiple Choice Reading Comprehension tests.

Table 1 shows the number of participants per edition. Table 2 shows the mean and best result of participants per edition based on accuracy, which was the main measure from 2003 to 2008, and the secondary measure from 2009.

**Table 1** Statistics about QA at CLEF campaign over the years

|                | Number of questions | Languages | Participants | Submitted runs | Monolingual runs | Cross-lingual runs |
|----------------|---------------------|-----------|--------------|----------------|------------------|--------------------|
| <b>Era I</b>   |                     |           |              |                |                  |                    |
| 2003           | 200                 | 3         | 8            | 17             | 6                | 11                 |
| 2004           | 200                 | 7         | 18           | 48             | 20               | 28                 |
| 2005           | 200                 | 8         | 24           | 67             | 43               | 24                 |
| 2006           | 200                 | 9         | 30           | 77             | 42               | 35                 |
| <b>Era II</b>  |                     |           |              |                |                  |                    |
| 2007           | 200                 | 10        | 22           | 37             | 20               | 17                 |
| 2008           | 200                 | 11        | 21           | 51             | 31               | 20                 |
| <b>Era III</b> |                     |           |              |                |                  |                    |
| 2009           | 500                 | 9         | 11           | 28             | 26               | 2                  |
| 2010           | 200                 | 7         | 13           | 49             | 45               | 4                  |
| <b>Era IV</b>  |                     |           |              |                |                  |                    |
| 2011           | 120                 | 5         | 12           | 62             | 62               | 0                  |
| 2012           | 160                 | 7         | 11           | 43             | 40               | 3                  |
| 2013           | 284                 | 5         | 11           | 54             | 54               | 0                  |
| 2014           | 56                  | 5         | 4            | 29             | 29               | 0                  |
| 2015           | 89                  | 6         | 5            | 18             | 18               | 0                  |
| Total          | 2609                | –         | 190          | 580            | 436              | 144                |



**Table 2** Results at QA@CLEF based on accuracy

|                | Monolingual |         | Multilingual |            |
|----------------|-------------|---------|--------------|------------|
|                | Mean        | Best    | Mean         | Best       |
| <b>Era I</b>   |             |         |              |            |
| 2003           | 0.29        | 0.49 IT | 0.17         | 0.45 IT-EN |
| 2004           | 0.24        | 0.46 NL | 0.15         | 0.35 EN-NL |
| 2005           | 0.29        | 0.65 PT | 0.18         | 0.40 EN-FR |
| 2006           | 0.28        | 0.68 FR | 0.25         | 0.49 PT-FR |
| <b>Era II</b>  |             |         |              |            |
| 2007           | 0.23        | 0.54 FR | 0.11         | 0.42 EN-FR |
| 2008           | 0.24        | 0.64 PT | 0.13         | 0.19 RO-EN |
| <b>Era III</b> |             |         |              |            |
| 2009           | 0.41        | 0.61 EN | 0.16         | 0.18 EU-EN |
| 2010           | 0.51        | 0.72EN  | 0.28         | 0.30 EN-RO |
| <b>Era IV</b>  |             |         |              |            |
| 2011           | 0.16        | 0.48 EN | –            | –          |
| 2012           | 0.26        | 0.65 EN | 0.29         | 0.29 RO-EN |
| 2013           | 0.26        | 0.49 EN | –            | –          |
| 2014           | 0.26        | 0.59 FR | –            | –          |
| 2015           | 0.31        | 0.58 EN | –            | –          |

## 2 Era I: 2003–2006. Ungrouped Mainly Factoid Questions Asked Against Monolingual Newspapers; Exact Answers Returned

### 2.1 Description

The introduction of multilinguality represented not only a great novelty in the QA research field, but also a good chance to stimulate the QA community to develop and evaluate multilingual systems.

In 2003 (Magnini et al. 2004), three languages were addressed in the monolingual tasks (Dutch, Italian and Spanish), while in the bilingual tasks questions were formulated in five source languages (Dutch, French, German, Italian and Spanish) and answers were searched in an English document collection.

In 2003, the task consisted of returning automatically (i.e. with no manual intervention), a ranked list of [docid, answer] pairs per question such that the retrieved document supported the answer. Participants were given 200 questions for each language sub-task, and were allowed to submit up to three responses per query. They were asked to retrieve either a 50-byte snippet of text extracted from the document collections, which provided exactly the amount of information required, or an exact answer. Each returned run consisted of either entirely 50-byte answers or exact answers, but not a mixture. Twenty questions had no known answer in the target corpora: systems indicated their confidence that there was no answer in the document collection by returning “NIL” instead of the [docid,

answer] pair. There was general agreement about their usefulness in assessing the systems' performances, so a certain number of NIL questions were created in all QA campaigns until 2008. In the first year of the track, only Factoid questions were considered, i.e. fact-based questions, asking for the name of a person, a location, the extent of something, the day on which something happened, etc. Participants were not required to return a supporting context for their answer until 2006.

Participants had one week to process the questions. Since no manual intervention of any kind was allowed, we asked participants to freeze their systems before downloading the queries from our QA@CLEF website. Before the start of the evaluation exercise, we released detailed guidelines with the necessary information about the required format of the submissions. We also put online a checking routine with which participants could make sure that their responses were in compliance with the guidelines.

In 2004 the QA@CLEF track attracted considerable attention within the CLEF framework (Magnini et al. 2005). It involved the main QA task, a Spanish pilot task and iCLEF, the interactive track. The main task included more European languages than CLEF 2003 and all the cross-language combinations between them were exploited to set up a number of different subtasks.

In 2004, the main task was repeated without changes but for the addition of four new languages, and two new question types: Definition and a new answer type for Factoid, namely Manner. Definition questions asked for the position of a person (e.g. *Who is Tony Blair?*), the meaning of an acronym (e.g. *What is UNICEF?*) or giving information about something (e.g. *What is the atom?*).

Despite the demand for radical innovation, a conservative approach was also preferred in 2005 (Vallin et al. 2006), as the procedures consolidated in the last two campaigns seemed to need further investigation before moving to the next stage. Although the task remained basically the same as that of 2004, some minor changes were made: the question types Manner and Object were discontinued and, at the same time, the concept of Temporal Restriction was introduced. This was the property of restricting answers to a given question (of any type) to those that were valid only when associated with an event, when occurring on a particular date, or when taking place within a time interval. Temporal restrictions were used in a subset of CLEF questions in all years up until 2008.

In 2006 (Magnini et al. 2007), the most significant innovation was the introduction of List questions, which had also been considered for previous competitions, but had been avoided due to the issues that their selection and assessment implied. In contrast to TREC, where each answer was listed as a separate, self contained response to the question, at CLEF the list was contained within a single response; this means that the answer was found in one passage of the document set that spelled out the entire list. Under this aspect, these single response List questions did not differ from a traditional Factoid question. Moreover, such questions could require either "closed lists" as answers, consisting in a number of specified items, or "open lists", where an unspecified number of correct answers could be returned. In case of closed lists, correct partial answers, where only some of the expected items were present, were evaluated as inexact. This kind of question was introduced in order

to allow a multilingual investigation of List questions without requiring a separate evaluation procedure.

Other important innovations of the 2006 campaign were the possibility to return up to ten exact answers per question, and the requirement to additionally provide up to ten text snippets, i.e. substrings of the specified documents giving the actual context of the exact answer in order to justify it.

## 2.2 Results

Each single answer was judged by human assessors, who assigned to each response a unique label: either right, wrong, unsupported or inexact. Assessors were told to judge the submissions from a potential user's point of view, because the evaluation should take into consideration the future portability of QA systems. They analyzed both the answers themselves and the context-, i.e. the document that supported the answer-, in which they appeared.

Answers were judged to be incorrect (W) when the answer-string did not contain the answer or when the answer was not responsive. In contrast, a response was considered to be correct (R) when the answer-string consisted of nothing more than the exact, minimal answer (or contained the correct answer within the 50 bytes long string) and when the document returned supported the response. Unsupported answers (U) were correct but it was impossible to infer that they were responsive from the retrieved document. Answers were judged as non-exact (X) when the answer was correct and supported by the document, but the answer string missed bits of the response or contained more than just the exact answer. Answers to definition questions were judged considering their usefulness for a potential user who was assumed to know nothing of the person or the organization addressed by the question.

The main evaluation measure was accuracy (the proportion of correct answers). Additional measures were applied to offer secondary results from other perspectives. Such measures were Confidence Weighted Score (CWS) (Voorhees 2002), K and K1 (Herrera et al. 2005). Moreover, the performance over NIL questions was measured using precision, recall and the F-measure (harmonic mean).

In the first era (2003–2006), monolingual factoid QA showed a steady improvement, starting at 49% of correct answers in the first year and increasing to 68% in the fourth (2006). Interestingly, the best system was for a different language in each of those years: Italian, Dutch, Portuguese and French respectively. The improvement can be accounted for by the adoption of increasingly sophisticated techniques gleaned from other monolingual tasks at TREC and NTCIR, as well as at CLEF. However, during the same time, cross-lingual QA showed very little improvement, remaining in the range of 35–49% of correct answers. The bottleneck for cross-lingual QA is Machine Translation and clearly the required improvement in MT systems had not been realised by participants in the task.

As a general remark, systems that attempted a cross-language task in addition to a monolingual one did not show a similar performance trend in the two tasks, the cross-language task recording much lower scores. For example, the QRISTAL system developed by Synapse Développement in 2005 (Laurent et al. 2007) participated in four tasks having French as target language (namely monolingual French, English-French, Italian-French, and Portuguese-French). While it obtained good results in the monolingual task, reaching 64%, its performance decreased in the cross-language tasks, scoring 39.50, 25.50, 36.50% respectively. Another example is the 2006 Priberam system (Cassan et al. 2007): it performed well in the monolingual Portuguese task, with an accuracy of 69%, but in crosslingual Spanish-Portuguese task its accuracy dropped to 29%. Similarly, the system scored 51% in the monolingual Spanish task, but only 34.4% in the cross-lingual Portuguese-Spanish task.

Regarding the type of questions, systems obtained the best results over definition questions. The best example was represented by Montes-y-Gómez et al. (2006), who obtained an 80% accuracy over definition questions by using patterns for creating from the source documents a database with definitions to persons and organizations.

Participant systems were based on pipeline architectures. These architectures relied on IR modules using keywords from questions for recovering candidate documents. So, these systems used to fail when questions contained a different rewording of candidate documents.

### **3 Era II: 2007–2008. Grouped Questions Asked Against Newspapers and Wikipedias; Exact Answers Returned**

#### **3.1 Description**

In 2007 (Giampiccolo et al. 2008), the questions were grouped into clusters, each of which referred to the same topic. This meant that co-reference could be used between entities mentioned in the cluster of questions. In these cases, the supporting document for the second answer could be not the same as that for the first answer. Another major novelty for 2007 concerned the documents. Up to 2006, each data collection comprised a set of newspaper articles provided by ELRA/ELDA. Then, in 2007, Wikipedia dated 2006 was used as well, capitalising on the experience of the WiQA pilot task (Jijkoun and de Rijke 2007). Thus, for example, the answer to a question in French could be found in a French newspaper article (as in previous years), in a French Wikipedia entry, or both. One of the main reasons for using the Wikipedia collections was to make a first step towards Web-formatted corpora; as a huge amount of information was available on the Web, this was considered a desirable next level in the evolution of QA systems.

The 2007 task proved to be much more difficult than expected because of the grouped questions. Not only did groups include co-reference but, in addition,

the questions became intrinsically more complicated because they were no longer semantically self-contained, as the simple factoids of earlier campaigns had been. Instead, they effectively developed a theme cumulatively. In order to allow participants more time to further study this problem, the exercise was repeated almost without changes in 2008 (Forner et al. 2009).

Again, participant systems relied on IR modules for finding correct answers, using coreference for completing questions in the same cluster.

## 3.2 Results

In the second era (2007–2008), the task became considerably more difficult because questions were grouped around topics and in particular because, sometimes, it was necessary to use coreference information across different questions. Monolingual performance dropped 14%, from its previous high of 68% in 2006 to 54% in 2007, and then increased to 64% in 2008. At the same time, crosslingual performance decreased from the 2006 figure of 49% (PT-FR) in the previous Era to 42% (EN-FR) in 2007. Relative to the change in monolingual system performance, this was a smaller decrease. Then, in 2008, the figure fell to 19%. This dramatic change can be explained by the fact that the monolingual systems in Era II were roughly the same as those in Era I.

## 4 Era III: 2009–2010. Ungrouped Questions Against Multilingual Parallel-Aligned EU Legislative Documents; Passages or Exact Answers Returned

### 4.1 Description

By 2005, we realized that there was an upper bound of 60% of accuracy in system performance, despite more than 80% of the questions being answered by at least one participant. We understood that we had a problem of error propagation in the traditional QA pipeline (Question Analysis, Retrieval, Answer Extraction, Answer Selection/Validation). Thus, in 2006 we proposed a task called Answer Validation Exercise (AVE) (Peñas et al. 2007). The aim was to produce a change in QA architectures to give more responsibility to the validation step. In AVE we assumed there was a previous step of hypothesis generation and the hard work had to be done in the validation step. This is a kind of classification task that could take advantage of Machine Learning. The same idea is behind the architecture of IBM's Watson (DeepQA project) that successfully participated in Jeopardy (Ferrucci et al. 2010).

After the three editions of AVE (described in the pilot task Section) we tried to transfer our conclusions to the main QA task at CLEF 2009 and 2010 (Peñas et al.

2010a,b). The first step was to introduce the option of leaving questions unanswered. This is an easy way of testing systems' confidence: if a system is not sure about its answers, it can decide to leave unanswered a question instead of risking an incorrect answer. This is related to the development of validation technologies. Then, we needed a measure able to reward systems that reduce the number of questions answered incorrectly without affecting system accuracy, by leaving unanswered the questions they estimated they couldn't answer. The measure was an extension of accuracy called  $c@1$  (Peñas and Rodrigo 2011), which was the main measure at QA@CLEF from 2009.  $c@1$  adds value to the traditional accuracy depending on the performance answering questions. Thus, if a system obtains a high performance answering questions, it receives a high reward when leaving unanswered a question.

Until 2009, the target collections consisted of newspaper articles, which were comparable but not parallel and, as a consequence, the answer might be present in more than one language collection, even though not in all. On the contrary, in 2009 and 2010 campaigns a parallel aligned corpus was used, which made the task completely multilingual, i.e. questions had an answer in all target languages.

The 2009 evaluation track, called ResPubliQA, represented a radical change with respect to the previous QA campaigns at CLEF. The exercise was aimed at retrieving answers to a set of 500 questions. The required output was not an exact answer but an entire paragraph, and the collection JRC-Acquis was from a specific domain, i.e. European legislation. Moreover, three new question types were introduced, in an attempt to move away from the factoid paradigm: Procedure, Purpose and Reason. Finally, the choice of a specific domain represented a first step towards the definition of a realistic user model. The issue of identifying potential users of QA systems had been a matter of discussion among the track organizers for a long time, but in the campaigns held so far, the focus was on proposing a general task in order to allow systems to perfect the existing techniques. In 2009, time seemed ripe to make the task more realistic and introduce a user model. While looking for a suitable context, improving the efficacy of legal searches in the real world seemed an approachable field, as the retrieval of information from legal texts was an issue of increasing importance given the vast amount of data which had become available in electronic form in the previous years.

The design of the ResPubliQA 2010 evaluation campaign was to a large extent a repetition of the previous year's exercise. However, this year participants had the opportunity to return both paragraph and exact answers as system output. Another novelty was the addition of a portion of the EuroParl collection which contained transcribed speeches from the European Parliament. Moreover, Reason and Purpose questions, which had been found to be too similar to one another, were duly merged into one category, Reason-Purpose. At the same time, two new question types were introduced, Other and Opinion. In the case of the latter, it was thought that speeches within EuroParl might express interesting opinions.

## 4.2 Results

In the third era (2009–2010), the task changed to one of paragraph retrieval while at the same time the questions and document collection became more difficult. Monolingual performance started at a similar level of 61% in 2009 and then rose to 72% in 2010. Cross lingual performance was 18% (EU-EN) in 2009 and rose to 30% (EN-RO) in 2010. These very low figures can be accounted for by the fact that there was very little participation in the cross-lingual task during the third era.

However, this change was not enough. Almost all systems continued relying on IR engines to retrieve relevant passages and then trying to extract the exact answer from them. This is not the change in the architecture we expected, and again, results did not go beyond the 60% pipeline upper bound. Finally, we understood that the change in the architecture means putting more effort into the development of answer validation/selection technologies. For this reason, the task was reformulated and the step of retrieval was put aside for a while, focusing on the development of technologies able to work with a single document, and to answer questions about it.

## 5 Era IV: 2011–2015. Questions About a Single Document Using a Related Document Collection as Background Information. Multiple Choice Reading Comprehension Tests

### 5.1 Description

In the 2011 formulation of the task (Peñas et al. 2011), the step of retrieval was put aside for a while, focusing on the development of technologies able to work with a single document, and to answer questions about it.

In the new setting, we started again decomposing the problem into hypothesis generation and validation. Thus, in the QA4MRE task we tested systems only for the validation step. Together with the questions, the organization provided a set of candidate answers. This gave the evaluation the format of traditional Multiple Choice Reading Comprehension tests.

This development parallels the introduction in 2009 of the Machine Reading Program (MRP) by DARPA in the USA. The goals of the program were to develop systems that perform deep reading of small numbers of texts in given domains and to answer questions about them. Analogously to QA4MRE, the MRP program involved batteries of questions for the evaluation of system understanding. However, testing queries were structured according to target ontologies, forcing participant teams to focus on the problem of document transformation into the formal representation defined by these target ontologies. Thus the Machine Reading challenge had to pass through the Information Extraction paradigm. In QA4MRE we followed a different approach leaving the door open to find synergies with emerging research

areas such as those related to Distributional Semantics, Knowledge Acquisition, and Ontology Induction. For this reason, we were agnostic with respect to the query language and the machine internal representation. Thus, questions and answers were posed in natural language.

The QA4MRE task focused on the reading of single documents and the identification of the answers to a set of questions. Questions were in the form of multiple choice, each having several options, and only one correct answer. The detection of correct answers might eventually require various kinds of inference and the consideration of previously acquired background knowledge from reference document collections. Although the additional knowledge obtained through the background collection may be used to assist with answering the questions, the answer had to be found among the facts contained in the given test documents. Thus, reading comprehension tests did not require only semantic understanding but they assumed a reasoning process that involves using implications and presuppositions, retrieving the stored information and performing inferences to make information explicit. Many different forms of knowledge took part in this process: linguistic, procedural, world and common sense knowledge. All these forms coalesce during processing and it is sometimes difficult to clearly distinguish and reconstruct them in a system that needs additional knowledge and inference rules in order to understand the text and to give sensible answers.

By giving only a single document per test, systems were required to understand every statement and to form connections across statements in case the answer was spread over more than one sentence. Systems were requested to (1) understand the test questions, (2) analyse the relation among entities contained in questions and entities expressed by the candidate answers, (3) understand the information contained in the documents, (4) extract useful pieces of knowledge from the background collections, and (5) select the correct answer from the five alternatives proposed.

In 2013, we ran a pilot task called Entrance Exams (Peñas et al. 2013). In all previous tasks, questions were posed by organizers with the aim of evaluating automatic systems under different reading abilities, types of questions, inference degree, etc. In the challenge of “Entrance Exams”, the goal was to test systems in a real scenario, like in a Turing test. Thus, systems were evaluated under the same conditions humans are evaluated to enter the University of Tokyo. For this purpose, some exercises about Reading Comprehension were extracted from actual exams. This exercise was organized in coordination with the “Entrance Exams” task at NTCIR. Exams were created by the Japanese National Center for University Admissions Tests and the “Entrance Exam” corpus was provided by NII’s Todai Robot Project and NTCIR.

In the 2014 and 2015, Entrance Exams was considered as the main task of the QA track.



## 5.2 Results

Average results were close to a 0.25 score of  $c@1$ , which is a value slightly higher than the random selection. Nevertheless, this average value is below the 0.5 score usually required to “pass” RC tests. These results showed that participant systems returned more incorrect answers than correct answers, which was not the expected behavior after reducing the importance of the IR component

Only one system from Synapse could give more correct than incorrect answers (Laurent et al. 2014, 2015; Laurent 2014). We realized that there were several issues such as the semantic gap between texts, questions and answers; external knowledge management; etc. Thus, the task should be directed to a simpler one with easier tests, as for example those for primary school as has been suggested in other studies (Clark and Etzioni 2016).

We detected that only a few systems left some questions unanswered. In these cases, despite the fact that some systems reduced considerably the amount of incorrect answers, only the best system in the 2013 edition could improve its overall  $c@1$  score.

## 6 Pilot Exercises

QA at CLEF was also an opportunity to experiment with several pilot tasks, whose common goal was to investigate how QA systems and technologies are able to cope with different types of questions from those proposed in the main task, experimenting with different scenarios. The following pilot tasks have been proposed over the years:

- Question Answering 2004 pilot task (Herrera et al. 2005): the task had a two-fold aim: (1) in the first place, the evaluation of Question Answering systems when they have to answer conjunctive lists, disjunctive lists and questions with temporal restrictions. (2) the evaluation of systems’ capability to give an accurate self-scoring about the confidence on their answers. Results of this pilot task were transferred to the main task by including temporal restrictions in questions and using two new measures,  $k$  and  $k1$ , as secondary measures.
- Real Time Question Answering (Noguera et al. 2007): the task proposed an exercise for the evaluation of QA systems within a time constraint, carried out in the 2006 campaign, and proposing new measures which combine Precision with the answer time. This task show the difficulty of answering some questions in a short period of time.
- Answer Validation Exercise (Peñas et al. 2007): the task consisted of a voluntary exercise to promote the development and evaluation of sub-systems aimed at validating the correctness of the answers given by a QA system. The basic idea was that once an [answer + snippet] pair is returned to a question by a QA system, an Answer Validation module has to decide whether the answer is correct

according to the supporting snippet. The results of this exercise stimulated the development of new QA systems and suggested a change in the main QA task in 2007.

- Question Answering over Speech Transcripts (Lamel et al. 2008): the aim of the task was to evaluate QA technology in a real multilingual speech scenario in which written and oral questions (factual and definitional) in different languages were formulated against a set of audio recordings related to speech events in those languages. The scenario was the European Parliament sessions in English, Spanish and French.
- Word Sense Disambiguation for Question Answering (Forner et al. 2009): this consisted of a pilot task that provided the questions and collections with already disambiguated word senses in order to study their contribution to QA performances.
- Question Answering using Wikipedia (Jijkoun and de Rijke 2007): the purpose was to see how IR and NLP techniques could be effectively used to help readers and authors of Wikipedia pages to access information spread throughout Wikipedia rather than stored locally on the pages. Specifically, the task involved detecting whether a snippet contained new information or whether it duplicated what was already known.
- GikiCLEF (Santos and Cabral 2010): following the previous GikiP pilot at GeoCLEF 2008, the task focused on open list questions over Wikipedia that require geographic reasoning, complex information extraction, and cross-lingual processing, for Bulgarian, Dutch, English, German, Italian, Norwegian, Portuguese, Romanian and Spanish.
- Processing Modality and Negation for Machine Reading (Morante and Daelemans 2011): this task was aimed at evaluating whether systems were able to understand extra propositional aspects of meaning like modality and negation. Modality is a grammatical category that expresses aspects related to the attitude of the speaker towards his/her statements, including certainty, factuality, and evidentially. Negation is a grammatical category that allows changing the truth value of a proposition. Modality and negation interact to express extra-propositional aspects of meaning. This task exploited the same topics and background collections of the Main Task. However, test documents were specifically selected to ensure the properties required for the questions. Participating systems had to decide whether given events in the texts were Asserted, Negated, or Speculated. The task was offered in English only in 2011 and 2012. In 2013 we integrated modality and negation into the Main Task by including some questions that required this kind of processing in order to answer correctly.
- Machine Reading on Biomedical Texts about Alzheimer's disease (Morante et al. 2013): this pilot task explored the ability of a system to answer questions using scientific language. The test posed questions in the Biomedical domain with a special focus on one disease, namely Alzheimer's. Texts were taken from PubMed Central related to Alzheimer's and from 66,222 Medline abstracts. Here, the specific domain enabled us to explore Machine Reading linked to controlled

vocabularies, entity types, and a predefined set of relations among these entity types. Thus, the task aimed at finding contact points with approaches based on Information Extraction.

## 7 Other Question Answering Tasks in CLEF

Two more QA tasks run together with QA track at CLEF during 2013, 2014 and 2015: Question Answering over Linked Data (QALD) and BioAsq.

### 7.1 *Question Answering Over Linked Data*

QALD is a series of evaluation campaigns on multilingual question answering over linked data, with a strong emphasis on interlinked datasets and hybrid approaches using information from both structured and unstructured data. The key challenge lies in translating the users' information needs into a form such that they can be evaluated using standard Semantic Web query processing and inference techniques (Lopez et al. 2013).

The core task of QALD aims at all question answering systems that mediate between a user, expressing his or her information need in natural language, and semantic data. The participating systems had to return either the correct answers, or a SPARQL query that retrieves these answers.

QALD acknowledge also that a lot of further information is still available only in textual form, both on the web and in the form of labels and abstracts in linked data sources. This is why QALD proposed also a subtask focused on the integration of both structured and unstructured information in order to gather answers. Given a version of DBpedia, containing both RDF data and free text available in the DBpedia abstracts, and a natural language question or keywords, participating systems had to retrieve the correct answer(s).

### 7.2 *BioAsq*

BioASQ aimed at assessing (Tsatsaronis et al. 2015):

- large-scale classification of biomedical documents onto ontology concepts (semantic indexing),
- classification of biomedical questions onto relevant concepts,
- retrieval of relevant document snippets, concepts and knowledge base triples,
- delivery of the retrieved information in a concise and user-understandable form.

The challenge comprised two tasks: (1) a large-scale semantic indexing task and (2) a question answering task.

**BioASQ 1, Large-Scale Semantic Indexing:** the goal was to classify documents from the PubMed digital library into concepts of the MeSH2 hierarchy. Here, new PubMed articles that had not been annotated yet were collected on a weekly basis.

These articles were used as test sets for the evaluation of the participating systems. As soon as the annotations were available from the PubMed curators, the performance of each system was calculated by using standard information retrieval measures as well as hierarchical ones.

In order to provide an on-line and large-scale scenario, the task was divided into three independent batches. In each batch five test sets of biomedical articles were released consecutively. Each of these test sets was released on a weekly basis and the participants had 21 h to provide their answers.

**Task BioASQ 2, Biomedical Semantic Question Answering:** the goal of this task was to provide a large-scale question answering challenge where the systems should be able to cope with all the stages of a question answering task, including the retrieval of relevant concepts and articles, as well as the provision of natural language answers.

It comprised two phases: In phase A, BioASQ released questions in English from benchmark datasets created by a group of biomedical experts. There were four types of questions: yes/no questions, factoid questions, list questions and summary questions. Participants had to respond with relevant concepts (from specific terminologies and ontologies), relevant articles (PubMed and PubMedCentral articles), relevant snippets extracted from the relevant articles and relevant RDF triples (from specific ontologies).

In phase B, the released questions contained the correct answers for the required elements (concepts, articles, snippets and RDF triples) of the first phase. The participants had to answer with exact answers as well as with paragraph-sized summaries in natural language (dubbed ideal answers).

The task was split into five independent batches. The two phases for each batch were run with a time gap of 24 h. For each phase, the participants had 24 h to submit their answers. The evaluation in phase B was carried out manually by biomedical experts on the ideal answers provided by the systems.

## 8 Lessons Learnt

Over the years of CLEF QA campaigns, some lessons attached to the goals of each particular challenge have been learned. From the very beginning in 2003, the track had a strong focus on multilinguality and tried to promote the development of translanguing systems. Despite all the efforts made in this direction (translating questions in many different languages and using comparable and parallel corpora) systems targeting different languages cannot be strictly compared and no definite

conclusions can be drawn. Nevertheless, the resources developed allow the comparison of the same system across different languages, which is very important for QA developers that work in several languages, as the performances of different systems targeting the same language can be assessed comparatively.

The final methodology, regarding multilinguality, was implemented in 2009 and 2010, where both questions and documents had parallel translations. Thus, the systems that participated in several languages served as reference points for comparison across languages.

Another lesson learned concerned how the evaluation setting determines the participant systems architecture. By 2005, it became clear that there was an upper bound of 60% of accuracy in systems' performance, although more than 80% of the questions were answered by at least one participant. It emerged that there was a problem of error propagation in the most used QA pipeline (Question Analysis, Retrieval, Answer Extraction, Answer Selection/Validation). Thus, in 2006 a pilot task called Answer Validation Exercise (AVE) was proposed, aimed at fostering a change in QA architectures by giving more relevance to the validation step. In AVE, the assumption was that after a preliminary step of hypothesis over-generation, the validation step decides whether the candidate answer is correct or not. This is a kind of classification task that could take advantage of Machine Learning. The same idea is behind the architecture of IBM's Watson (DeepQA project) that successfully participated at Jeopardy.

After the three campaigns of AVE an attempt was made to transfer the conclusions to the QA main task at CLEF 2009 and 2010. The first step was to introduce the option of leaving questions unanswered, which was related to the development of validation technologies necessary to develop better QA systems. A suitable measure was also needed in order to reward systems that reduced the number of questions answered incorrectly without affecting system accuracy, by leaving unanswered those questions whose answers the system was not confident about. The measure was an extension of accuracy called  $c@1$ , tested during 2009 and 2010 QA campaigns at CLEF, and used also in subsequent evaluations

However, this was not the change in the architecture that was expected, as almost all systems continued using indexing techniques to retrieve relevant passages and tried to extract the exact answer from that. Moreover, results did not go beyond the 60% pipeline upper bound.

Therefore, the conclusion was that, in order to foster a real change in the QA system architecture, a previous development of answer validation/selection technologies was required. For this reason, the new formulation of the task after 2010 left the retrieval step aside to focus on the development of technologies able to work with a single document, answering questions about it and using the reference collections as sources of background knowledge that help the answering process.

Another lesson learned was that most participants reduced the concept of answer validation simply to the task of answer ranking. For this purpose, they developed similarity based approaches that did not decide whether there is a correct answer or not among candidates. Generally, they simply trusted the ranking score to exceed a given threshold. So, going back to the question of whether systems achieved

sufficient performance to ensure that there will be a qualitative difference when trying full QA scenario, the answer is: possibly not.

Over the years, it has become clear that groups working on Question Answering are not making use of background knowledge collections very much. At most, systems might locate some possibly relevant material from the background collection through simple matching, and then use associated information to help rank the potential answers. Tying in with the point above on answer ranking, indicates the difficulty of introducing inference/reasoning into processing.

Regarding the construction of background collections, we have learned it is very difficult to adequately define Background Knowledge, and to specify the types and sources that must be considered to solve the full QA scenario. There are increasingly more sources of linked / relational data that, potentially, can be used. However, language goes beyond a predefined set of relations among entities and values. That was the reason to propose the use of text collections inviting participants to acquire propositional knowledge useful for textual inferences. We have not obtained much of value in this regard.

Despite the difficulty of defining Background Knowledge, we have learned that if we want to use text collections to contextualize system readings, we must be very careful not to introduce any kind of bias. Therefore, the idea of creating a background collection able to contextualize a single text can be formulated as a classical Information Retrieval task, i.e. to retrieve all relevant documents and only them. Any methodological approach must take this ideal as reference and try to approximate it as much as possible.

The last editions showed also some open-challenges for QA systems. On one hand, QA systems still find difficulties when dealing with complex questions, where different pieces of information must be gathered from several sources, as for example documents and knowledge bases (Harabagiu et al. 2006). On the other hand, there were poor results over questions where correct answers appear with different wording in documents and questions. This rewording is common in tests oriented to assess document understanding. Moreover, these questions may ask about information that appears implicitly, but not explicitly, in texts. These challenges should receive a higher attention for current evaluations if the community wants to improve QA systems.

**Acknowledgements** This work has been partially funded by the Spanish Research Agency (Agencia Estatal de Investigación) LIHLITH project (PCIN-2017-085/AEI).

## References

- Cassan A, Figueira H, Martins A, Mendes A, Mendes P, Pinto C, Vidal D (2007) Priberam's question answering system in a cross-language environment. In: Peters C, Clough P, Gey FC, Karlgren J, Magnini B, Oard DW, de Rijke M, Stempfhuber M (eds) Evaluation of multilingual and multi-modal information retrieval : seventh workshop of the cross-language evaluation forum (CLEF 2006). Revised selected papers. Lecture notes in computer science (LNCS) 4730. Springer, Heidelberg, pp 300–309

- Clark P, Etzioni O (2016) My computer is an honor student - but how intelligent is it? Standardized tests as a measure of AI. *AI Mag* 37(1):5–12
- Ferrucci DA, Brown EW, Chu-Carroll J, Fan J, Gondek D, Kalyanpur A, Lally A, Murdock JW, Nyberg E, Prager JM, Schlaefer N, Welty CA (2010) Building Watson: an overview of the DeepQA project. *AI Mag* 31(3):59–79
- Forner P, Peñas A, Agirre E, Alegria I, Forascu C, Moreau N, Osenova P, Prokopidis P, Rocha P, Sacaleanu B, Sutcliffe RFE, Sang EFTK (2009) Overview of the Clef 2008 multilingual question answering track. In: Peters C, Deselaers T, Ferro N, Gonzalo J, Jones GJF, Kurimo M, Mandl T, Peñas A (eds) *Evaluating systems for multilingual and multimodal information access: ninth workshop of the cross-language evaluation forum (CLEF 2008)*. Revised selected papers. *Lecture notes in computer science (LNCS)* 5706. Springer, Heidelberg, pp 262–295
- Giampiccolo D, Forner P, Herrera J, Peñas A, Ayache C, Forascu C, Jijkoun V, Osenova P, Rocha P, Sacaleanu B, Sutcliffe RFE (2008) Overview of the CLEF 2007 multilingual question answering track. In: Peters C, Jijkoun V, Mandl T, Müller H, Oard DW, Peñas A, Petras V, Santos D (eds) *Advances in multilingual and multimodal information retrieval: eighth workshop of the cross-language evaluation forum (CLEF 2007)*. Revised selected papers. *Lecture notes in computer science (LNCS)* 5152. Springer, Heidelberg, pp 200–236
- Harabagiu S, Lacatusu F, Hickl A (2006) Answering complex questions with random walk models. In: *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '06*, pp 220–227
- Herrera J, Peñas A, Verdejo F (2005) Question answering pilot task at CLEF 2004. In: Peters C, Clough P, Gonzalo J, Jones GJF, Kluck M, Magnini B (eds) *Multilingual information access for text, speech and images: fifth workshop of the cross-language evaluation forum (CLEF 2004) revised selected papers*. *Lecture notes in computer science (LNCS)* 3491. Springer, Heidelberg, pp 581–590
- Jijkoun V, de Rijke M (2007) Overview of the wiqua task at CLEF 2006. In: Peters C, Clough P, Gey FC, Karlgren J, Magnini B, Oard DW, de Rijke M, Stempfhuber M (eds) *Evaluation of multilingual and multi-modal information retrieval : seventh workshop of the cross-language evaluation forum (CLEF 2006)*. Revised selected papers. *Lecture notes in computer science (LNCS)* 4730. Springer, Heidelberg, pp 265–274
- Lamel L, Rosset S, Ayache C, Mostefa D, Turmo J, Comas P (2008) Question answering on speech transcriptions: the QAST evaluation in CLEF. In: *Proceedings of the international conference on language resources and evaluation, LREC 2008, 26 May–1 June 2008, Marrakech, Morocco*
- Laurent D (2014) English run of synapse développement at entrance exams 2014. In: Cappellato L, Ferro N, Halvey M, Kraaij W (eds) *CLEF 2014 labs and workshops, notebook papers*. *CEUR workshop proceedings (CEUR-WS.org)*. ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>, pp 1404–1414
- Laurent D, Séguéla P, Nègre S (2007) Cross lingual question answering using QRISTAL for CLEF 2006. In: Peters C, Clough P, Gey FC, Karlgren J, Magnini B, Oard DW, de Rijke M, Stempfhuber M (eds) *Evaluation of multilingual and multi-modal information retrieval : seventh workshop of the cross-language evaluation forum (CLEF 2006)*. Revised selected papers. *Lecture notes in computer science (LNCS)* 4730. Springer, Heidelberg, pp 339–350
- Laurent D, Chardon B, Nègre S (2014) French run of synapse développement at entrance exams 2014. In: Cappellato L, Ferro N, Halvey M, Kraaij W (eds) *CLEF 2014 labs and workshops, notebook papers*. *CEUR workshop proceedings (CEUR-WS.org)*. ISSN 1613-0073. <http://ceur-ws.org/Vol-1180/>, pp 1415–1426
- Laurent D, Chardon B, Nègre S, Pradel C, Séguéla P (2015) Reading comprehension at entrance exams 2015. In: Cappellato L, Ferro N, Jones GJF, SanJuan E (eds) *CLEF 2015 labs and workshops, notebook papers*. *CEUR workshop proceedings (CEUR-WS.org)*. ISSN 1613-0073. <http://ceur-ws.org/Vol-1391/>
- Lopez V, Unger C, Cimiano P, Motta E (2013) Evaluating question answering over linked data. *Web Semantics: Sci Serv Agents World Wide Web* 21(0):3–13. Special Issue on Evaluation of Semantic Technologies

- Magnini B, Romagnoli S, Vallin A, Herrera J, Peñas A, Peinado V, Verdejo MF, de Rijke M (2004) The multiple language question answering track at CLEF 2003. In: Peters C, Braschler M, Gonzalo J, Kluck M (eds) *Comparative evaluation of multilingual information access systems: fourth workshop of the cross-language evaluation forum (CLEF 2003) revised selected papers*. Lecture notes in computer science (LNCS) 3237. Springer, Heidelberg, pp 471–486
- Magnini B, Vallin A, Ayache C, Erbach G, Peñas A, de Rijke M, Rocha P, Simov KI, Sutcliffe RFE (2005) Overview of the CLEF 2004 multilingual question answering track. In: Peters C, Clough P, Gonzalo J, Jones GJF, Kluck M, Magnini B (eds) *Multilingual information access for text, speech and images: fifth workshop of the cross-language evaluation forum (CLEF 2004) Revised selected papers*. Lecture notes in computer science (LNCS) 3491. Springer, Heidelberg, pp 371–391
- Magnini B, Giampiccolo D, Forner P, Ayache C, Jijkoun V, Osenova P, Peñas A, Rocha P, Sacaleanu B, Sutcliffe RFE (2007) Overview of the CLEF 2006 multilingual question answering track. In: Peters C, Clough P, Gey FC, Karlgren J, Magnini B, Oard DW, de Rijke M, Stempfhuber M (eds) *Evaluation of multilingual and multi-modal information retrieval : seventh workshop of the cross-language evaluation forum (CLEF 2006)*. Revised selected papers. Lecture notes in computer science (LNCS) 4730. Springer, Heidelberg, pp 223–256
- Montes-y-Gómez M, Pineda LV, Pérez-Coutiño MA, Soriano JMG, Arnal ES, Rosso P (2006) A full data-driven system for multiple language question answering. In: Peters C, Gey FC, Gonzalo J, Jones GJF, Kluck M, Magnini B, Müller H, de Rijke M (eds) *Accessing multilingual information repositories: sixth workshop of the cross-language evaluation forum (CLEF 2005)*. Revised selected papers. Lecture notes in computer science (LNCS) 4022. Springer, Heidelberg, pp 420–428
- Morante R, Daelemans W (2011) Overview of the QA4MRE pilot task: annotating modality and negation for a machine reading evaluation. In: Petras V, Forner P, Clough P, Ferro N (eds) *CLEF 2011 working notes*. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1177/>
- Morante R, Krallinger M, Valencia A, Daelemans W (2013) Machine reading of biomedical texts about alzheimer’s disease. In: Forner P, Navigli R, Tufis D, Ferro N (eds) *CLEF 2013 working notes*. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Noguera E, Llopis F, Ferrández A, Escapa A (2007) Evaluation of open-domain question answering systems within a time constraint. In: 21st International conference on advanced information networking and applications (AINA 2007). Workshops proceedings, vol 1, May 21–23, 2007, Niagara Falls, Canada, pp 260–265
- Peñas A, Rodrigo A (2011) A simple measure to assess non-response. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies - vol 1*. Association for Computational Linguistics, HLT ’11, pp 1415–1424
- Peñas A, Rodrigo Á, Sama V, Verdejo F (2007) Overview of the answer validation exercise 2006. In: Peters C, Clough P, Gey FC, Karlgren J, Magnini B, Oard DW, de Rijke M, Stempfhuber M (eds) *Evaluation of multilingual and multi-modal information retrieval : seventh workshop of the cross-language evaluation forum (CLEF 2006)*. Revised selected papers. Lecture notes in computer science (LNCS) 4730. Springer, Heidelberg, pp 257–264
- Peñas A, Forner P, Rodrigo A, Sutcliffe RFE, Forascu C, Mota C (2010a) Overview of ResPubliQA 2010: question answering evaluation over European legislation. In: Braschler M, Harman DK, Pianta E, Ferro N (eds) *CLEF 2010 working notes*. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073, <http://ceur-ws.org/Vol-1176/>
- Peñas A, Forner P, Sutcliffe RFE, Rodrigo A, Forascu C, Alegria I, Giampiccolo D, Moreau N, Osenova P (2010b) Overview of ResPubliQA 2009: question answering evaluation over European legislation. In: Peters C, Di Nunzio GM, Kurimo M, Mandl T, Mostefa D, Peñas A, Roda G (eds) *Multilingual information access evaluation vol. I text retrieval experiments – tenth workshop of the cross-language evaluation forum (CLEF 2009)*. Revised selected papers. Lecture notes in computer science (LNCS) 6241. Springer, Heidelberg, pp 174–196



- Peñas A, Hovy EH, Forner P, Rodrigo A, Sutcliffe RFE, Forascu C, Sporleder C (2011) Overview of QA4MRE at CLEF 2011: question answering for machine reading evaluation. In: Petras V, Forner P, Clough P, Ferro N (eds) CLEF 2011 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1177/>
- Peñas A, Miyao Y, Hovy E, Forner P, Kando N (2013) Overview of QA4MRE 2013 entrance exams task. In: Forner P, Navigli R, Tufis D, Ferro N (eds) CLEF 2013 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Santos D, Cabral LM (2010) GikiCLEF: expectations and lessons learned. In: Peters C, Di Nunzio GM, Kurimo M, Mandl T, Mostefa D, Peñas A, Roda G (eds) Multilingual information access evaluation vol. I text retrieval experiments – tenth workshop of the cross-language evaluation forum (CLEF 2009). Revised selected papers. Lecture notes in computer science (LNCS) 6241, Springer, Heidelberg, pp 212–222
- Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, Weissenborn D, Krithara A, Petridis S, Polychronopoulos D, Almirantis Y, Pavlopoulos J, Baskiotis N, Gallinari P, Artières T, Ngonga A, Heino N, Gaussier É, Barrio-Alvers L, Schroeder M, Androutsopoulos I, Paliouras G (2015) An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinf* 16:138:1–138:28
- Vallin A, Magnini B, Giampiccolo D, Aunimo L, Ayache C, Osenova P, Peñas A, de Rijke M, Sacaleanu B, Santos D, Sutcliffe RFE (2006) Overview of the CLEF 2005 multilingual question answering track. In: Peters C, Gey FC, Gonzalo J, Jones GJF, Kluck M, Magnini B, Müller H, de Rijke M (eds) Accessing multilingual information repositories: sixth workshop of the cross-language evaluation forum (CLEF 2005). Revised selected papers. Lecture notes in computer science (LNCS) 4022. Springer, Heidelberg, pp 307–331
- Voorhees EM (2000) Overview of the TREC-9 question answering track. In: Proceedings of the ninth text retrieval conference, TREC 2000, Gaithersburg, Maryland, USA, November 13–16, 2000
- Voorhees EM (2002) Overview of TREC 2002 question answering track. In: Voorhees EM, Buckland LP (eds) Proceedings of the eleventh text retrieval conference (TREC 2002). NIST Publication 500-251, pp 57–68
- Voorhees EM, Tice DM (1999) The TREC-8 question answering track evaluation. In: Text retrieval conference TREC-8, pp 83–105

# Evolution of the PAN Lab on Digital Text Forensics



**Paolo Rosso, Martin Potthast, Benno Stein, Efstathios Stamatatos, Francisco Rangel, and Walter Daelemans**

**Abstract** PAN is a networking initiative for digital text forensics, where researchers and practitioners study technologies for text analysis with regard to originality, authorship, and trustworthiness. The practical importance of such technologies is obvious for law enforcement, cyber-security, and marketing, yet the general public needs to be aware of their capabilities as well to make informed decisions about them. This is particularly true since almost all of these technologies are still in their infancy, and active research is required to push them forward. Hence PAN focuses on the evaluation of selected tasks from the digital text forensics in order to develop large-scale, standardized benchmarks, and to assess the state of the art. In this chapter we present the evolution of three shared tasks: plagiarism detection, author identification, and author profiling.

---

P. Rosso (✉)

PRHLT Research Center, Universitat Politècnica de València, Valencia, Spain  
e-mail: [proso@dsic.upv.es](mailto:proso@dsic.upv.es)

M. Potthast

Text Mining and Retrieval, Leipzig University, Leipzig, Germany  
e-mail: [martin.pothast@uni-leipzig.de](mailto:martin.pothast@uni-leipzig.de)

B. Stein

Web Technology and Information Systems, Bauhaus-Universität Weimar, Weimar, Germany  
e-mail: [benno.stein@uni-weimar.de](mailto:benno.stein@uni-weimar.de)

E. Stamatatos

Dept. of Information and Communication Systems Engineering, University of the Aegean, Samos, Greece  
e-mail: [stamatatos@aegean.gr](mailto:stamatatos@aegean.gr)

F. Rangel

Autoritas Consulting S.A, Valencia, Spain

PRHLT Research Center, Universitat Politècnica de València, Valencia, Spain  
e-mail: [francisco.rangel@autoritas.es](mailto:francisco.rangel@autoritas.es)

W. Daelemans

CLiPS - Computational Linguistics Group, University of Antwerp, Antwerp, Belgium  
e-mail: [walter.daelemans@uantwerpen.be](mailto:walter.daelemans@uantwerpen.be)

© Springer Nature Switzerland AG 2019

N. Ferro, C. Peters (eds.), *Information Retrieval Evaluation in a Changing World*, The Information Retrieval Series 41,  
[https://doi.org/10.1007/978-3-030-22948-1\\_19](https://doi.org/10.1007/978-3-030-22948-1_19)

# 1 Introduction

PAN<sup>1</sup> has become one of the main events for the digital text forensics community and it gathers a large audience of experts from information retrieval, natural language processing, and machine learning. The first two editions of PAN were organized in the form of workshops (2007–2008) at the conferences SIGIR 2007 and ECAI 2008 respectively. Since 2009, shared tasks have been organized at PAN, since 2010 Labs at CLEF, and since 2011 also at FIRE. At CLEF we have organized 31 shared tasks on authorship, originality, and trust: plagiarism detection (2010–2015), author identification (2011–2017), author profiling (2013–2017), Wikipedia vandalism detection (2010–2011), Wikipedia quality flaw detection (2012), sexual predator identification (2012), and author obfuscation (2016–2017). Each shared task had a considerable impact on its respective research field. Table 1 overviews key figures of the PAN Lab at CLEF in terms of registrations, runs/software, notebooks, attendees, and followers (Gollub et al. 2013; Potthast et al. 2014a; Stamatatos et al. 2015b; Rosso et al. 2016; Potthast et al. 2017). Since 2012 all of our shared tasks invite participants for *software* submissions instead of run submissions: more than 300 pieces of software have been submitted to PAN 2012 through PAN 2017, which have been repeatedly evaluated using the TIRA experimentation platform (Gollub et al. 2012a,b).

At FIRE<sup>2</sup> we organized 10 PAN shared tasks on text reuse/plagiarism detection in several languages (Arabic, Gujarati, Hindi, Persian) (Barrón-Cedeno et al. 2013; Gupta et al. 2012, 2013; Bensalem et al. 2015; Asghari et al. 2016), on source code texts (Flores et al. 2014, 2015), as well as on author profiling (Bengali, Hindi, Kannada, Malayalam, Russian,<sup>3</sup> Tamil and Telegu<sup>4</sup>) also addressing novel research aspects such as personality recognition in source code (Rangel et al. 2016a).

In this chapter we will describe three of the shared tasks that we have organized at CLEF: plagiarism detection, author identification, and author profiling. The rest of this chapter is structured as follows. The next section is devoted to plagiarism detection: evolution of tasks, evaluation framework, and submitted approaches. Section 3 is on the evolution of tasks in author identification (closed/open set attribution, verification, clustering, diarization, and style breach detection) and the submitted approaches. Section 4 is on author profiling and its evolution (age, gender, personality, and language variety), background about the employed corpora, and the performance of the submitted approaches. The last section contains conclusions and discusses research aspects that we plan to address in the near future in the framework of the PAN Lab at CLEF.

---

<sup>1</sup>Initially, PAN stood for “Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection” <http://pan.webis.de>.

<sup>2</sup>At CLEF 2010 in Padua, Carol Peters suggested cross-fertilization across evaluation forums.

<sup>3</sup><http://en.rusprofilinglab.ru/rusprofiling-at-pan>.

<sup>4</sup><http://nlp.amrita.edu:8080/INLI/Test.html>.

**Table 1** Key figures of the PAN shared tasks at CLEF

|               | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
|---------------|------|------|------|------|------|------|------|------|
| Followers     | 151  | 181  | 232  | 286  | 302  | 333  | 337  | 347  |
| Registrations | 53   | 52   | 68   | 110  | 103  | 148  | 143  | 191  |
| Runs/software | 27   | 27   | 48   | 58   | 57   | 54   | 37   | 34   |
| Notebooks     | 22   | 22   | 34   | 47   | 36   | 52   | 29   | 30   |
| Attendees     | 25   | 36   | 61   | 58   | 44   | 74   | 40   | 48   |

## 2 Plagiarism Detection

The first two editions of PAN were organized as workshops at the SIGIR 2007 and ECAI 2008 conferences. With the third edition at the SEPLN 2009 conference, PAN was organized for the first time as a (single) shared task: plagiarism detection. As it turned out, the time was ripe for this task, also evidenced by the comparably high number of first-time participants for a single shared task at that time.

Regardless of the fact that research on plagiarism detection was lacking (from algorithmic and conceptual perspectives) in various respects, the most pressing deficit probably was the missing evaluation and comparison of existing approaches. For decades, scientists published their findings individually, making up their own evaluation data, methodology, and performance measures ad hoc—often without consulting the relevant literature first. Comparisons between different related approaches were hardly ever conducted, so that an interested researcher entering the field had the problem of guessing which of the approaches reflected the state of the art or provided the strongest baseline. This shortcoming was addressed by our series of shared tasks: key contributions include corpora that have been created manually via crowdsourcing or semi-automatically, the implementation of a sophisticated evaluation setup including custom-built search engines, large-scale manual essay writing to simulate text reuse, and the first-time definition of suitable performance measures, incorporating the specifics of the task.

We want to point out that a technology that claims to detect plagiarism, in fact, does not. Instead, it detects evidence of text reuse, which may or may not be sufficient to judge whether an author plagiarized with a certain probability. While our task should have been called *text reuse detection* instead of plagiarism detection, we recognized that the misnomer was still justified: people both within and outside of academia search for “plagiarism detection”, whereas hardly anyone is familiar with the term “text reuse detection”. In light of this fact, we continue to use the former term for our task to ensure that interested people will find it in the future, while focusing our attention on text reuse detection as well as mentioning the connection in appropriate places.

From a public relations perspective the timing of this topic could not have been better. By 2011, when the shared task was in its third edition, we were fully prepared for the major plagiarism scandal that hit Germany in that year: it was

discovered that the then-minister of defense, Karl Theodor zu Guttenberg, had plagiarized considerable parts of his dissertation, copying bits and pieces of text from more than 200 sources to fill up his 400-page thesis, resulting in both the loss of his doctor's degree and his position. With the ensuing public outcry a number of other theses of famous politicians were checked and dozens more cases were found. These realistic cases of plagiarism provided us with an intriguing baseline to judge whether plagiarism detection technology was mature enough to detect such cases, as well as renewed the research interest in the task itself, which lasts to this day. Interestingly, dedicated plagiarism detection software hardly played a role in resolving these cases; the analyses were done manually, by up to hundreds of people who collaborated within Wikis to crowdsource their detection efforts. Also note in this regard, that there is a high chance that the collection of real cases is skewed towards ease of detection, while the difficult cases where plagiarizing authors carefully paraphrased the text they reused may have gone unnoticed. Aside from privacy issues this is another reason why these real cases can only serve as an additional source of data for evaluating plagiarism detectors. New corpora are needed to render the task of analysis more realistic beyond the detection of verbatim copy-paste operations.

## 2.1 *Evolution of Tasks*

The plagiarism detection task was organized for seven successive years, starting in 2009. In previous research (Stein et al. 2007), we interpreted plagiarism detection as a two-step retrieval process, which, given a suspicious document, consists of the tasks: (1) a source retrieval task, executed against a large collection of reference documents such as the web, followed by (2) a text alignment task, performed on the retrieved candidate sources against the suspicious documents with the objective of extracting plagiarized passages.<sup>5</sup>

In the first edition of our task, called external plagiarism detection, our goals were twofold (Potthast et al. 2009): (1) to create the first benchmark for plagiarism detection under the aforementioned retrieval process, consisting of suspicious documents with and without plagiarism, the former being drawn from a large-scale reference collection of documents obtained from the Project Gutenberg<sup>6</sup>; (2) to scale that setup to an—at that time—large-enough size so that participants would not just compare all pairs of documents to each other but to force them to do some sort of source retrieval within their approach. To ensure that the extraction

---

<sup>5</sup>In the beginning, the two tasks were called “candidate retrieval” and “detailed comparison” respectively. Later on, as the importance of evaluating these tasks in isolation became clear, we found our initial choice of names to be too unspecific and decided to rename them for clarification as “source retrieval” and “text alignment”.

<sup>6</sup><http://www.gutenberg.org>.

of plagiarized passages from pairs of suspicious documents and retrieved candidate sources would be non-trivial, we applied so-called obfuscation strategies in order to emulate plagiarists attempting to hide their plagiarism by paraphrasing the reused texts. We implemented a number of automatic obfuscation strategies, which, for lack of a working paraphrasing model, ranged from random text operations to parts-of-speech-preserving word reorderings. Although the automatic obfuscation strategies served as a good baseline for a bag-of-words-oriented plagiarism detector, the obfuscated passages obtained were still unreadable and hence lacked an appropriate semantics. To render the obfuscation step more realistic, we resorted to crowdsourcing the required paraphrases on Amazon's Mechanical Turk, which was still a rather new tool at the time. The resulting paraphrases were manually written, so that they served as a more realistic sample of human paraphrasing ability at passage level; still, the obfuscated passages were inserted at random into suspicious documents and could be spotted rather easily by human readers. Nevertheless, it turned out we were among the first to use crowdsourcing for paraphrasing acquisition, and the first to do so at passage level, so that we published a corresponding spin-off corpus, the Webis Crowd Paraphrase Corpus (Webis-CPC-11). In addition, we also provided an in-depth analysis of the corpus quality as well as machine learning technology that allowed for automatic quality assessment during paraphrase acquisition, severely reducing construction costs (Burrows et al. 2013). As became clear upon the review of the 14 approaches submitted, none of the participants actually implemented source retrieval but all of them went to great lengths to compare every document from the reference collection to each of the suspicious documents. In hindsight, the number of 41,000 documents was already too small to impose a source retrieval step. Significantly increasing the corpus size was still impossible for us since we had already exhausted the entire Project Gutenberg for our purposes. And, simply adding documents from a different source (and hence: different genre) would have been too easy to be recognized and undone. As a consequence, instead of treating plagiarism detection as an atomic task, we decided to evaluate source retrieval and text alignment in isolation. Within the next two iterations of the plagiarism detection task (Potthast et al. 2010a, 2011), the evaluation setup was refined with a focus on text alignment, while we started to build a new and independent evaluation setup specifically suited to source retrieval.

In addition to the task above, we invented and hosted the task of *intrinsic* plagiarism detection (Stein et al. 2011). The goal of this task is to identify plagiarized passages without exploiting an external document collection. I.e., tackling this task means finding evidence for writing style changes, which in turn may indicate that some text from another author has been copied into a suspicious document at hand. Although rather clear in its design, intrinsic plagiarism detection contains a number of considerable challenges; it was the foray of PAN into the field of writing style analysis and can be seen as the precursor of various authorship-related tasks that PAN hosts today. Similar to the external plagiarism detection task, the task was repeated three times in a row, refining its setup from year to year.

Starting in 2012 (Potthast et al. 2012a), a new evaluation setup for plagiarism detection was ready for use. This setup enabled (and still enables) us to

evaluate source retrieval tasks in much more realistic settings, separating it from the evaluation of text alignment tasks. The setup was used for four successive years (Potthast et al. 2013a, 2014b, 2015; Hagen et al. 2015). While the text alignment task did not change much, for the source retrieval task a new search engine called ChatNoir (Potthast et al. 2012b) was built, which indexed the entire ClueWeb 2009 (ClueWeb09 2009). Using this search engine, we compiled—via expert crowdsourcing—also a new corpus of manually created plagiarism. In particular, more than 20 writers were recruited, where each writer was asked to write essays about some topic of her choice from the TREC ad hoc track, yielding a total of 300 essays. Each essay was supposed to be of 5000 words length, and the research required to write the essay had to be conducted with ChatNoir’s web interface, reusing text from the web pages found. Moreover, the writers were instructed to obfuscate the reused text passages in a way they deemed sufficient to successfully pass plagiarism detectors. Some writers spent significant effort to do so while others did not, resulting in a range of case difficulties. This corpus, called the Webis Text Reuse Corpus 2012 (Webis-TRC-12) (Potthast et al. 2013b), formed the basis for several spin-off research inquiries (e.g., analyzing the writing behavior of writers during search (Hagen et al. 2016)) as well as follow-up shared tasks on author diarization at PAN (Stamatatos et al. 2016; Tschuggnall et al. 2017). Participants of the task had to treat a given essay from the corpus as suspicious document and to use the ChatNoir API to retrieve all sources from which an essay’s author reused text fragments. The queries and downloads of potential sources of the submitted approaches were meticulously logged to measure their performance in terms of retrieval effort and recall. To relieve participants from the task of also implementing text alignment technology, a “source oracle” was provided, which classified a downloaded document either as true or as false source.

## 2.2 *Evaluation Framework*

The evaluation framework that has been developed within the series of shared tasks on plagiarism detection had a strong impact on the community (Potthast et al. 2010b). It is employed to this day and helps to evaluate new algorithms, ensuring the comparability of new and historical evaluation results. The evaluation framework consists of three components: (1) a collection of corpora for text alignment and source retrieval, (2) a static, reproducible web search environment for source retrieval, and (3) tailored performance measures for both tasks.

Altogether 26 corpora have been constructed for our shared tasks. The collection includes corpora that were submitted to shared tasks that specifically invited their participants to submit not just software, but also data. Following the example of our shared tasks, spin-offs have been organized in subsequent years at other conferences, dedicated to specific languages not previously covered. Our corpus collection serves as a diverse resource, allowing for the evaluation of plagiarism detectors under

many different scenarios, especially regarding the difficulty of the to-be-detected plagiarism cases.

The static web search environment is comprised of the web search engine ChatNoir, which indexes the ClueWeb 2009, the ClueWeb 2012, and (as of 2017) the CommonCrawl, delivering search results in milliseconds while using a state-of-the-art retrieval model and a standard user interface. The web search environment comes along with a framework that allows for browsing web pages from the aforementioned corpora as if being in the live web, while serving clicks on hyperlinks with the version of the crawled page instead of the live version. The user behavior in the framework can be logged, allowing for reproducible, large-scale user studies, as well as evaluating various search-based approaches, including source retrieval.

Our text alignment measures consider the *granularity* of a detection result (i.e., they penalize a detector if it returns bits and pieces of a plagiarized passages instead of the reused passage as a whole) as well as formulas for precision and recall that discount multiple, overlapping detections for a given suspicious document. The measures can be combined into the so-called “PlagDet score”, which allows for an absolute ranking among evaluated plagiarism detectors (Potthast et al. 2010b). The source retrieval measures are based on recall but consider the effort in terms of queries and downloads as well. Again, multiple detections of the same source documents are discounted when retrieving web pages that are duplicates of each other.

### 2.3 Submitted Approaches

Over the years, many approaches haven been submitted to the plagiarism detection task and its variants—too many to review all of them here. Table 2 shows the distribution of participants across tasks. A total of 74 approaches have been submitted to external plagiarism detection and its successor task text alignment, 26 approaches have been submitted to source retrieval, and 10 to intrinsic plagiarism detection. The approaches submitted in or after 2012 for text alignment, and in or after 2013 for source retrieval, have been archived in an operational state and are still available for re-evaluation within TIRA.

The approaches submitted to a task showed certain commonalities—a fact, which allowed us to discern and organize a general (task-specific) retrieval process, comprising a number of task-specific steps. Each step in turn can be operationalized in numerous ways, and, once the algorithmic pattern was revealed to participants, it guided their developments and allowed newcomers to catch up quickly without having to reinvent the wheel.

A text alignment approach generally is divided into three steps: (1) seeding, (2) extension, and (3) filtering. The names of these steps are borrowed from gene sequence alignment, a task in bioinformatics that relates to text alignment in that the problem structure is similar, albeit not the solution space. The seeding step takes as



**Table 2** An overview of author identification tasks at PAN evaluation campaigns

| Year | Tasks                          | Language                      | Submissions |
|------|--------------------------------|-------------------------------|-------------|
| 2009 | External plagiarism detection, | English, German, Spanish      | 10          |
|      | Intrinsic plagiarism detection | English                       | 4           |
| 2010 | External plagiarism detection, | English, German, Spanish      | 18          |
|      | Intrinsic plagiarism detection | English                       | 2           |
| 2011 | External plagiarism detection, | English                       | 9           |
|      | Intrinsic plagiarism detection | English                       | 4           |
| 2012 | Text alignment                 | English                       | 10          |
|      | Source retrieval               | English                       | 5           |
| 2013 | Text alignment                 | English                       | 9           |
|      | Source retrieval               | English                       | 9           |
| 2014 | Text alignment                 | English                       | 11          |
|      | Source retrieval               | English                       | 6           |
| 2015 | Text alignment data submission | English, Farsi, Urdu, Chinese | 8           |
|      | Source retrieval               | English                       | 5           |
|      | External plagiarism detection  | Arabic                        | 3           |
|      | Intrinsic plagiarism detection | Arabic                        | 2           |
| 2016 | Text alignment                 | Persian                       | 9           |
|      | Text alignment data submission | Persian                       | 5           |
|      | Source retrieval               | English                       | 1           |
| 2017 | Text alignment                 | Russian                       | 1           |
|      | Source retrieval               | English                       | 1           |

input two documents and outputs matches between them in terms of pairs of phrases (one from each document) for which a matching heuristic checks similarity in order to argue about equivalent semantics. A commonly used matching heuristic outputs all matching word 4-grams whose words have been synonym-normalized, stemmed, and sorted alphabetically. The heuristic thereby raises the matching probability of two word 4-grams, even if the author of a plagiarized document paraphrases a text resulting in new word ordering at phrase level. Another heuristic employs so-called stop word 8-grams, which are 8-grams consisting only of the stop words in order of appearance in a text (Stamatatos 2011). Since plagiarists often focus on exchanging content words rather than function words, matching sequences of stop words are a telltale sign of reused text. The finesse of devising matching heuristics between two texts determines to a great extent how well a text alignment approach works, since matches that were not identified during seeding render the subsequent step of extension difficult if not impossible. In this regard, the more matching heuristics are employed simultaneously, the better. The extension step takes as input the matches obtained from seeding, and outputs the boundaries of pairs of passages (one from each document) that may have been copied and pasted by the original author before paraphrasing. When interpreting each match as a point in a two-dimensional plane spanned by the two documents' characters, clustering technology can be used to extend text regions with dense amounts of seeds towards larger passages for which a human would judge that they have obviously been reused in bulk. Finally, the filtering step implements some postprocessing to exclude results that seem

implausible according to certain criteria; most participants, however, just remove detections that would otherwise harm their performance in terms of granularity.

A source retrieval approach is divided into four steps: (1) chunking, (2) keyphrase extraction, (3) query formulation, and (4) query and download scheduling. The chunking step takes as input a suspicious document and outputs possibly overlapping chunks that cover the text. Many chunking strategies have been devised, but it turned out that non-overlapping 150-word chunks are sufficient. Each chunk is used as input for the keyphrase extraction step, where the  $k$  keywords or phrases that best describe the contents of a chunk (e.g., according to a  $tf \cdot idf$  ranking) are returned. The small chunk size renders the task of selecting the top  $k$  for  $k < 20$  easier. In addition, keyphrases from the entire document may be extracted to allow for querying the document's general topic. The query formulation step takes a set of keyphrases as input and returns queries consisting of at most five phrases, formulated by combining individual phrases and often started from the top-ranked keyphrase. In the query and download scheduling step, the queries are submitted to a search engine while trying to ensure that the most promising queries are submitted first, and the most promising search results are downloaded first to minimize the time to result. Here, queries comprising nouns were found to be most successful. The number of downloads per query may vary dependent on whether one wants to maximize the  $F$ -measure or just the recall. In the latter case, downloading more search results yields significant returns, whereas the likelihood of finding a second true positive detection after the first one in a given search result is small. I.e., the next query scheduled should be used after a true positive detection, whereas one may explore up to a hundred search results per query. The examples given for the aforementioned steps are, in fact, the ones followed by the most effective approach in terms of recall (0.89), dwarfing the best previously achieved recall (0.59) (Hagen et al. 2017). A number of additional heuristics render this approach comparable in terms of its effort to the previously best one while maintaining its recall.

### 3 Author Identification

Author identification aims at revealing the authors behind texts. It is an active research area (Stamatatos 2009) associated with important applications mainly in the humanities (e.g., unmasking the authors of novels published anonymously or under aliases), forensics (e.g., identifying the author of harassing messages, linking proclamations of terrorist groups), and social media analytics (e.g., revealing multiple user accounts controlled by the same person, verifying the authenticity of posts). Author identification tasks can be either supervised (i.e., the training texts are labeled with authorship information) (Argamon and Juola 2011; Juola and Stamatatos 2013) or unsupervised (i.e., authorship information is either not available or not reliable) (Stamatatos et al. 2016; Tschuggnall et al. 2017).

What makes author identification challenging is that it deals with the personal style of authors. In contrast to other factors, like topic or sentiment, usually

style is not associated with certain words and there is no consensus about its quantification. Moreover, it is especially hard to discover style markers (i.e., style-related quantifiable textual features) that remain unaffected in topic shifts or genre variations. Another crucial factor is text-length. For very long documents (e.g., novels), there are quite reliable methods (Koppel et al. 2007). However, when short or very short (e.g., tweets) texts are considered, it is much harder to retain high effectiveness.

### 3.1 Evolution of Tasks

A significant part of PAN activities is related to author identification. PAN evaluation campaigns since 2011 explored several tasks as summarized in Table 3 and described below:

- Closed-set attribution: Given a set of candidate authors and some texts unquestionably written by each one of them, the task is to find the most likely author among them for another text of disputed authorship.
- Open-set attribution: This is similar to the previous task. However, it is possible that none of the candidate authors is the author of the disputed text.
- Verification: Given a set of texts all written by the same author, the task is to examine whether another text is also written by that author.
- Clustering: Given a set of texts of unknown authorship, the task is to group them by authorship.

**Table 3** An overview of author identification tasks at PAN evaluation campaigns

| Year | Tasks  | Genre   | Language                          | Submissions |
|------|--|---|-----------------------------------|-------------|
| 2011 | Closed-set attribution,<br>Open-set attribution,<br>Verification | Emails  | English                           | 16          |
| 2012 | Closed-set attribution,<br>Open-set attribution,<br>Clustering   | Fiction   | English                           | 12          |
| 2013 | Verification   | Textbooks, Fiction,<br>Newspaper articles       | English, Greek,<br>Spanish        | 18          |
| 2014 | Verification   | Essays, Reviews,<br>Newspaper articles, Fiction | Dutch, English,<br>Greek, Spanish | 13          |
| 2015 | Verification   | Essays, Reviews,<br>Newspaper articles, Fiction | Dutch, English,<br>Greek, Spanish | 18          |
| 2016 | Clustering,<br>Diarization                                       | Reviews,<br>Newspaper articles                  | Dutch, English,<br>Greek          | 10          |
| 2017 | Clustering,<br>Style breach detection                            | Reviews,<br>Newspaper articles                  | Dutch, English,<br>Greek          | 9           |

- **Diarization:** Given a text that may be written by multiple co-authors, the task is to identify the authorial components of each co-author.
- **Style breach detection:** Given a text that may be written by multiple co-authors, the task is to detect all borders where authors switch.

In the different editions over the years, variations of the main task are examined. For example, the closed-set attribution task in 2011 focused on a large pool of candidate authors while the 2012 edition examined a small set of candidate authors. The first editions of the author verification task assumed that all texts within a verification case are in the same thematic area and belong to the same genre while the 2015 edition examined more challenging cross-topic and cross-genre cases. The first edition of the clustering task (2016) considered full texts while the 2017 edition focused on paragraph-length texts.

It has to be underlined that, in most of the cases, previous work in these tasks was extremely limited (e.g., author verification, clustering, diarization). PAN campaigns attracted the attention of multiple research groups around the world and contributed to enrich the literature in those areas. For all of these tasks, new benchmark corpora were developed covering several natural languages and genres (as shown in Table 3) that became the standard in the field. Moreover, appropriate evaluation measures were proposed for each task taking into account both crisp answers and confidence scores (Stamatatos et al. 2014, 2016; Tschuggnall et al. 2017). Especially, for the author verification task, emphasis was given to the fact that some cases could be left unanswered since in the applications related with this task it is better not to give an answer rather than giving a wrong answer. It was also demonstrated that authorship clustering can also be seen as a retrieval problem (Stamatatos et al. 2016).

The first editions of PAN related to author identification (2011–2012) were quite ambitious attempting to explore multiple tasks simultaneously. It soon became apparent that it is much better if each task is examined separately and in consecutive campaigns so that research groups are more mature and can develop more sophisticated approaches. Another important conclusion was that it is better to study simple rather than complicated tasks. For example, author verification can be seen as a fundamental task in author identification since any other task can be transformed into a series of author verification cases. Focusing on author verification enables us to better estimate the state-of-the-art performance in this area since we have to worry about fewer parameters (e.g., the number of candidate authors and the distribution of texts over the authors are not so crucial factors in verification as they are in closed-set attribution). Another example of a complicated task is author diarization. This can be decomposed into simpler tasks like style breach detection and clustering of short texts (Tschuggnall et al. 2017).

Another important outcome of PAN campaigns was to highlight the fact that there are strong relationships between author identification tasks. For example, author verification relates not only to closed-set and open-set attribution but to clustering as well. The top-performing approach in author verification at PAN 2015 was also the winning method in the clustering task in PAN 2016 (Bagnall 2016). Moreover, as already mentioned, authorship clustering is a basic building block in

the author diarization task. The latter is also strongly related with the task of intrinsic plagiarism detection considered in the early editions of PAN (Potthast et al. 2009, 2010a, 2011).

### 3.2 Submitted Approaches

Most of the author identification tasks attracted a large number of participant teams from all around the world. The submitted methods explored several models regarding the extraction of stylometric measures from texts and the attribution model. This section reviews the most important novelties and conclusions that can be drawn.

In the first editions of author identification tasks at PAN, it seemed that approaches based on a rich set of stylometric features combining several kinds of measures, including measures extracted by natural language processing tools (NLP), are the most promising ones (Argamon and Juola 2011). However, in subsequent shared tasks most of the top-performing submissions were based on low-level features like character and word n-grams. Such simplistic and language-independent features when combined with sophisticated attribution models can provide very good results (Stamatatos et al. 2014, 2015a, 2016; Tschuggnall et al. 2017). A particularly interesting and very effective approach is to apply neural network language models in stylometry as demonstrated by the character-level recurrent neural network model that won top-ranked overall positions in PAN 2015 and PAN 2016 tasks (Bagnall 2015, 2016). On the other hand, more sophisticated approaches exclusively based on syntactic analysis of texts by NLP tools were easily outperformed by simpler approaches. This can also be attributed to the fact that in most of the cases the NLP tools used by PAN participants were not specifically trained to handle the types of texts included in PAN corpora, therefore they provided quite noisy stylometric measures.

Certainly, the widest variety of methods submitted to PAN tasks refers to author verification. Table 4 shows the distribution of PAN participants per year according to several factors. Extrinsic verification models attempt to transform an author verification case from a one-class classification task to a binary classification task

**Table 4** Distribution of PAN participants in the author verification task

| Verification model | PAN 2013 | PAN 2014 | PAN 2015 |
|--------------------|----------|----------|----------|
| Intrinsic          | 13       | 10       | 11       |
| Extrinsic          | 3        | 3        | 7        |
| Eager              | 2        | 3        | 10       |
| Lazy               | 14       | 10       | 8        |
| Profile-based      | 4        | 1        | 4        |
| Instance-based     | 11       | 12       | 12       |
| Hybrid             | 1        | 0        | 2        |

by considering a collection of texts written by other authors (with respect to the author in question). A typical representative of this paradigm is the *Impostors* method introduced by Koppel and Winter (2014). Nevertheless, intrinsic verification models focus on one-class classification. In all three relevant editions of PAN, extrinsic verification models won top-ranked positions (Juola and Stamatatos 2013; Stamatatos et al. 2014, 2015a). However, the majority of PAN participants followed the intrinsic verification paradigm and only in the last edition of the shared task in PAN 2015 was there an increase of extrinsic models. A crucial open issue is how to find the most suitable set of external texts for a given verification case. The external documents used by relevant PAN submissions were downloaded from the World Wide Web with the help of a search engine and queries formed by texts of the training corpus (Seidman 2013; Khonji and Iraqi 2014).

Another important perspective is how to handle the training corpus. Eager methods attempt to build a binary classifier that learns to distinguish between positive (same-author) and negative (different-author) verification cases. Each verification case is an instance of this binary classification task and a classifier is trained based on the training corpus. Conversely, lazy methods essentially avoid extracting any general model from the training corpus and make their decisions separately for each evaluation case. The number of eager methods submitted to PAN increased over the years. This is certainly associated with the volume of the provided training corpora. In early editions of this task (PAN 2013) the training corpus consisted of a few dozens of verification cases while in the last two editions (PAN 2014 and PAN 2015) there were hundreds of verification cases in the training corpus. However, eager methods heavily depend on the representativeness of the training corpus. One eager method, trained on PAN 2014 corpora and among the best-performing submissions in PAN 2014 (Fréry et al. 2014), was also applied to PAN 2015 corpora (as a baseline model) and practically failed (Stamatatos et al. 2015a).

Verification models can also be described according to the way they handle texts of known authorship. One approach is to concatenate them and extract a single representation (profile-based paradigm). Another approach is to extract a separate representation from each known text (instance-based paradigm). Yet another case is to combine these two paradigms (hybrid methods). The majority of PAN submissions consistently follow the instance-based paradigm including the top-performing ones in most of the cases.

An important conclusion extracted from PAN shared tasks in author verification was that it is possible to combine different verification models and provide a robust approach with enhanced performance. A simple ensemble model that was based on averaging the answers of all PAN participants achieved better results than any of the individual models in the PAN 2013 and PAN 2014 author verification tasks. In the corresponding task at PAN 2015, the ensemble of all submissions was outperformed by some individual models mainly due to the relatively low average results of many participants. It is important to underline that the author verification task at PAN 2015 focused on very challenging cross-topic and cross-genre cases. However, the submission that ranked second-best overall was also based on a heterogeneous ensemble that combined several base verification models

(Moreau et al. 2015). Actually this approach was the most effective in the most challenging cross-genre corpus in Dutch (Stamatatos et al. 2015a). This clearly shows that heterogeneous ensembles is a promising approach and most suitable for challenging author verification tasks.

## 4 Author Profiling

Author profiling aims at identifying personal traits of an author on the basis of her/his writings. Traits such as gender, age, language variety, or personality are of high interest for areas such as marketing, forensics, or security. From the marketing viewpoint, to be able to identify personal traits from comments to blogs or reviews, may provide the companies with the possibility of better segmenting their audience, which is an important competitive advantage. From a forensic linguistics perspective one would like to be able to know the linguistic profile of the author of a harassing text message (language used by a certain type of people) and identify a certain type of person (language as evidence). From a security point of view, these technologies may allow to profile and identify possible delinquents or even terrorists. Traditional investigations in computational linguistics (Argamon et al. 2003) and social psychology (Pennebaker 2013) have been carried out mainly for English. Furthermore, pioneer researchers such as Argamon et al. (2003) or Holmes and Meyerhoff (2003), focused on formal and well-written texts. Although with the rise of social media, researchers such as Koppel et al. (2003) and Schler et al. (2006) have moved their focus to blogs and fora.

Since 2013 we have been organizing the author profiling task at PAN with several objectives. We have covered different profiling aspects (age, gender, language variety, personality), languages (Arabic, Dutch, English, Italian, Portuguese), and genres (blogs, reviews, social media, and Twitter). The international interest in the shared task is made evident by the number of participants from a large number of countries (Table 5). Furthermore, many have been researchers that have investigated further the performance of their approaches on the corpora that were developed for the shared task. For example, the best performing team in the three first editions used a second order representation which relates documents with author profiles and subprofiles (e.g., males talking about video games) (López-Monroy et al. 2015). The authors of Weren et al. (2014) investigated a high variety of different features on the PAN AP-2013 dataset and showed the contribution of information retrieval based features in age and gender identification. In this approach, the text to be identified was used as a query for a search engine. In Maharjan et al. (2014), the authors used MapReduce to approach the task with three million  $n$ -gram based features. They improved the accuracy as well as reduced the processing time considerably. Finally, the EmoGraph graph-based approach (Rangel and Rosso 2016) tried to capture how users convey verbal emotions in the morphosyntactic structure of the discourse. The sequence of grammatical categories is modeled as a graph which is enriched with topics, semantics of verbs, polarity, and emotions. They reported competitive results

**Table 5** An overview of author profiling tasks at PAN evaluation campaigns

| Year | Tasks                    | Genres                                  | Languages                              | Submissions |
|------|--------------------------|---|--|-------------|
| 2013 | Age, Gender              | Social media                            | English, Spanish                       | 21          |
| 2014 | Age, Gender              | Social media, Twitter<br>Blogs, Reviews | English, Spanish                       | 10          |
| 2015 | Age, Gender, Personality | Twitter                                 | English, Spanish<br>Italian, Dutch     | 22          |
| 2016 | Age, Gender              | Cross-genre                             | English, Spanish                       | 22          |
| 2017 | Gender, Language variety | Twitter                                 | English, Spanish<br>Arabic, Portuguese | 22          |

with the best performing systems at PAN 2013 and demonstrating its robustness against genres and languages at PAN 2014 (Rangel and Rosso 2015).

In the following sections we describe the evolution of the tasks, how the corpora have been built and the main approaches used by the participants, all from the perspective of the lessons learned during the organization of this task.

#### 4.1 Evolution of Tasks

In Table 5, a summary of the evolution of the author profiling tasks at PAN is shown.

The first edition was organized in 2013 (Rangel et al. 2013) with the aim of investigating the age and gender identification in a social media realistic scenario. We collected thousands of social media posts in English and Spanish with a high variety of topics. With respect to age, we considered three classes following what was previously done in Schler et al. (2006): 10s (13–17), 20s (23–27) and 30s (33–47). Furthermore, we wanted to test the robustness of the systems when dealing with fake age profiles such as sexual predators. Therefore, we included in the collection some texts from the previous year PAN shared task on sexual predator identification (Inches and Crestani 2012).

In the second edition (Rangel et al. 2014), we extended the task to other genres besides social media. Concretely, we focused also on Twitter, blogs, and hotel reviews, in English and Spanish. We realized the difficulty of obtaining quality labeled data and proposed a methodology to annotate age and gender. In 2014, we opted for modeling age in a continuous way and considered the following classes: 18–24; 25–34; 35–49; 50–64; 65+. Finally, the Twitter subcorpus was constructed in cooperation with RepLab (Amigó et al. 2014) in order to address also the reputational perspective (e.g., profiling social media influencers, journalists, professionals, celebrities, among others).

In 2015 (Rangel et al. 2015), besides the focus on age and gender identification, we introduced the task of personality recognition in Twitter. We maintained the age ranges defined in 2014 (except “50–64” and “65+” that were merged to “50–XX”) and, besides English and Spanish, we included also Dutch and Italian (only



gender and personality recognition). The objective of the shared task organized in 2016 (Rangel et al. 2016b) was to investigate the robustness of the systems in a cross-genre evaluation. That is, training the systems in one genre and testing its performance in other genres. Concretely, we provided Twitter data for training in English, Spanish, and Dutch. The approaches were then tested on blogs and social media genres in English and Spanish, and essays and reviews in Dutch.

Finally, in 2017 (Rangel et al. 2017) we introduced two novelties: the language variety identification (together with the gender), and the Arabic and Portuguese languages (besides English and Spanish). This is the first time a task has been organized covering together gender and language variety identification, and we obtained interesting insights relating both profiling aspects. Furthermore, we addressed language variety from fine-grain and course-grain perspective where varieties that are close geographically were grouped together (e.g. Canada and United States, Great Britain and Ireland, or New Zealand and Australia).

## 4.2 *Corpora Development*

The author profiling task organized at PAN has been focusing on social media texts. Our interest was to study how people use language in their daily lives. Thus, in 2013 we retrieved thousands of social media posts with a wide spectrum of topics. The ample diversity of topics made possible to go beyond standard cliches, for example, men writing about sports and women about shopping. Furthermore, people may use social media to talk about sex. Some users can also cross the line and commit sexual harassment. With the aim of investigating the robustness of the author profiling approaches in detecting possible predators, we included some texts from the previous year PAN task on sexual predator identification<sup>7</sup> (Inches and Crestani 2012). With this configuration, a realistic scenario was provided to the participants:

- A large dataset (big data).
- High variety of topics.
- Sexual conversations vs. sexual predators.
- Possible fake users and automatic generated content (e.g., chatbots).

This realistic scenario, however, presented some problems from the research perspective. The annotation (age and gender) was made on the basis of what the users self-reported, and they could have lied. Due to that, it was difficult to analyze errors: has the system failed or has it actually detected a fake profile? Therefore, we introduced a methodology to annotate data (and to not trust what users say). In the next subsections, we briefly describe this methodology for each trait.

---

<sup>7</sup>Texts from predators and adult-adult sexual conversations have been segmented into the corresponding age and gender groups.

### 4.2.1 Gender Annotation Based on Dictionary and Photos Review

Depending on the genre, the annotation of the gender was based on different methods. In the case of blogs or reviews, the starting point were lists of well-known users (e.g., celebrities or politicians on the one hand, colleagues or students on the other). Furthermore in case of Twitter, we took advantage of meta-information to label the profiles in two steps:

- Firstly, the user name was searched in a dictionary of proper nouns. Users with ambiguous names were discarded.
- Secondly, each profile photograph was visually reviewed in order to ensure the right gender. Users with ambiguous photography (e.g., non-personal photos) were discarded.

### 4.2.2 Age Annotation Based on LinkedIn Profiles

LinkedIn<sup>8</sup> is a professional network where people, among other things, can detail their resume. We looked for public LinkedIn profiles which share a personal blog URL or a Twitter account. We verified that the blog or the Twitter account existed, it was written in one of the languages we were interested in, and it was updated only by one person and this person was easily identifiable (we discarded organizational accounts). We looked for age information in the LinkedIn profile (in some cases the birth date is published). When this information was not available, we looked for the degree starting date in the education section. Following the information of Table 6, we figured out the age range. We discarded users whose education dates were not clear. To ensure the quality of the annotation, this process was done by two independent annotators and a third one decided in case of disagreement.

### 4.2.3 Personality Traits Annotation Based on BFI-10 Online Test

Personality may be defined along five traits using the Five Factor Theory (Costa and McCrae 2008), which is the most widely accepted model in psychology. The

**Table 6** Age range by degree starting date for data collected in the year 2014

| Degree starting date | Age group |
|----------------------|-----------|
| 2006-...             | 18-24     |
| 1997-2006            | 25-34     |
| 1982-1996            | 35-49     |
| 1967-1981            | 50-64     |
| ...-1966             | 65+       |

<sup>8</sup><https://www.linkedin.com>.

five traits are: openness to experience (O), conscientiousness (C), extraversion (E), agreeableness (A), and emotional stability/neuroticism (N). Personality traits, as well as users' gender and age, were self-assessed with the BFI-10 online test<sup>9</sup> (Rammstedt and John 2007) and reported as scores normalized between  $-0.5$  and  $+0.5$ .

The personality test consists of ten statements such as “I am a reserved person”, “I have few artistic interests”, or “I am sociable”. The user has to evaluate how much she/he agrees with each statement. Furthermore, she/he is asked for the age, gender, and Twitter account. This allowed us to retrieve the user's timeline and associate it with the profile aspects.

#### 4.2.4 Language Variety Annotation Based on Geographical Retrieval

A language variety is the specific form of a language that is shared by a group of people depending on their regional, social, or contextual situation. Taking advantage of Twitter geographical retrieval, we can obtain users who share a location and a language, and hence, a common language variety. To annotate users with their corresponding language variety, we have followed the following steps:

- Firstly, we decided which languages and language varieties will be part of the dataset. We selected four languages (Arabic, English, Portuguese and Spanish), and the varieties were selected following previous investigations (e.g. the selection of Arabic varieties followed (Sadat et al. 2014) as shown in Table 7).
- Varieties have been linked to geographical regions. For each language variety, the countries where this variety is used have been selected. Then, the capital cities (sometimes also the most populated cities) have been identified.
- Given the geographical coordinates of the capital cities, we have retrieved all the tweets generated in a radius around these coordinates (generally 15 km).
- Unique authors who wrote the retrieved tweets have been identified. Their entire timeline was then retrieved. Tweets written in other languages or retweets have been removed.
- Users whose tweets were not geotagged in the corresponding coordinates, or whose location did not coincide with the corresponding capital city have been removed. This avoids the inclusion of users who wrote when temporarily being in a particular place (e.g., tourists or temporary workers).

Although according to the Oxford English Dictionary the definition of dialect refers to “a variety of a language that is a characteristic of a particular group of the language's speakers” and “a language that is socially subordinated to a regional or national standard language”, the main criticism is that people from the same region are likely to talk about the same local topics. This may allow shallow topic-based

---

<sup>9</sup>We have created a web page with the BFI-10 test (<http://mypersonality.autoritas.net>) and promoted it in social media such as Twitter and Facebook.

**Table 7** Language varieties

| Arabic    | English       | Portuguese | Spanish   |
|-----------|---------------|------------|-----------|
| Egypt     | Australia     | Brazil     | Argentina |
| Gulf      | Canada        | Portugal   | Chile     |
| Levantine | Great Britain |            | Colombia  |
| Maghrebi  | Ireland       |            | Mexico    |
|           | New Zealand   |            | Peru      |
|           | United States |            | Spain     |
|           |               |            | Venezuela |

methods to achieve competitive results. However, the obtained results showed that the best results could not be achieved only with topic-based features since they did not capture other linguistic patterns that are even more common such as differences in used characters (e.g., in English *organise/organize*), parts-of-speech sequences (e.g., in Portuguese *quero quixar-me/quero-me queixar* *I want to complain*), or even words that appear only in some varieties (e.g., in Arabic, the words طارق (your remembrance), شعاد (but what about) and لقياك (meeting you) are only used in the Gulf variety.)

### 4.3 Submitted Approaches

Following Pennebaker investigations (Pennebaker 2013), most participants have combined different kinds of style-based features such as frequencies of punctuation marks, capital letters, quotations, and so on, together with Part-of-Speech tags or genre-specific features such as HTML-based features as image URLs, links, Twitter hashtags, or user mentions. Other stylistic markers such as the use of slang, contractions, or character flooding have been used as well.

Different content-based features have also been used: Latent Semantic Analysis, bag-of-words (weighted by frequency and tf-idf), dictionary-based words, topic-based words, entropy-based words, class-dependent words, named entities, etc. With respect to emotional features, some participants have extracted emotions, appraisal, admiration, positive/negative emoticons, positive/negative words, emojis, and sentiment words. Resources such as LIWC<sup>10</sup> have been widely used.

Language models based on different kinds of n-gram models (e.g., word, character) have been widely used in all the editions, obtaining competitive results, although almost always combined with other kinds of features. Other features such as readability indices (e.g., Flesch-Kinkaid, Gunning fog, SMOG, Coleman-Liau), information retrieval (the text to be identified was used as a query for a search engine), or collocations have been used by some participants. Finally, in recent

<sup>10</sup><https://liwc.wpengine.com>.

**Table 8** Best results at PAN

| Trait            | Arabic | Dutch  | English | Italian | Portuguese | Spanish |
|------------------|--------|--------|---------|---------|------------|---------|
| Age              | –      | –      | 0.8380  | –       | –          | 0.7955  |
| Gender           | 0.8031 | 0.9688 | 0.8592  | 0.8611  | 0.8700     | 0.9659  |
| Language variety | 0.8313 | –      | 0.8988  | –       | 0.9838     | 0.9621  |
| Personality      | –      | 0.0563 | 0.1442  | 0.1044  | –          | 0.1235  |

years, especially in 2017, deep learning approaches have been widely used, mainly based on distributed representations such as word and character embeddings.

With respect to classification algorithms and their evolution, most of the participants have approached the task with traditional machine learning algorithms such as Logistic Regression, Support Vector Machines, Naive Bayes, BayesNet, or Random Forest. There have also been participants who approached the task with distance-based methods. It is difficult to highlight the best algorithms due to the combination of them by participants, but in most cases the best performing teams used Support Vector Machines.

As previously said, deep learning methods have been widely used: Recurrent Neural Networks and Convolutional Neural Networks with configurations of attention mechanism, max-pooling layer, or fully-connected layer. Although these deep learning approaches obtained good results, they did not achieve the best ones.

In Table 8, best results at PAN per trait and language (accuracy for age, gender and language variety, RMSE for personality) were achieved in Twitter. Best results were obtained in 2015 in age, personality and gender in Dutch, English, Italian, and Spanish, in 2017 in language variety and gender identification in Arabic and Portuguese.

## 5 Conclusions

The shared tasks of PAN are designed both to measure the technical state of the art and to foster the development of new approaches for important problems in the field of digital text forensics. The shared task principle seems to be ideally suited for this endeavor; in particular, it attracts different research groups from different fields, which all have their own view and solution approach to tackle such kinds of “ill-posed” or fuzzy problems. The fuzziness of most of the PAN shared task problems has several causes: the complexity of language, the complexity of features to describe language phenomena, the complex distribution of the phenomena over text registers, or the missing theory about corpus size and robust feature quantification, to mention a few.

We, at PAN, address this challenging research situation by evolving our shared tasks. Stated differently, we are looking for the “right” question that we want to ask the research community. The three strands of task evolutions presented in this

chapter reflect this. However, the evolution must be driven carefully: we cannot completely re-model all tasks with each new PAN edition since (1) it may become too complicated for us to put together all pieces of the puzzle, and (2) we depend on the expertise that has been built up among the researchers of the participating teams, and we cannot require them to acquire and operationalize effective expertise from scratch each year. Hence we try to evolve the tasks in such a way that, on the one hand, they remain closely connected to the nature of the problem and, on the other hand, their variation brings enough insights to further develop the field. In this regard, we will continue the research on author identification, author profiling, or author obfuscation—although from different perspectives: cross-domain authorship attribution, style change detection, or multimodal author profiling (age and gender).

PAN has become a reference point in the digital text forensics community. Multiple shared tasks attracted a large number of participants and motivated research teams all over the world to start conducting research in this area. The corpora developed in the framework of PAN shared tasks have become standard benchmark datasets used in any subsequent study. Certainly, PAN corpora are far from ideal and sometimes they may suffer from low volumes of data, noise, or lack of realism. Therefore, maximizing the performance on those specific datasets should not be seen as panacea for the research community.

In addition, it is very important that PAN promotes reproducibility issues by requiring software submissions and encouraging participants to also provide their open-source code. All gathered approaches can be viewed as a library of tools, the largest in this area, available to replicate evaluation results and be applied to future corpora. The mere existence of this library enables the study of new tasks, like author obfuscation. PAN welcomes any other scientific use of this collection of software.

**Acknowledgements** The work of Paolo Rosso was partially funded by the Spanish MICINN under the research project MISMIS-FAKEHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31). The work on the author profiling data in Arabic was made possible by NPRP grant #9-175-1-033 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

## References

- Amigó E, Carrillo-de-Albornoz J, Chugur I, Corujo A, Gonzalo J, Meij E, de Rijke M, Spina D (2014) Overview of RepLab 2014: author profiling and reputation dimensions for online reputation management. In: Proceedings of the fifth international conference of the CLEF initiative
- Argamon S, Juola P (2011) Overview of the international authorship identification competition at PAN-2011. In: CLEF 2011 labs and workshop, notebook papers, 19–22 Sept 2011, Amsterdam, The Netherlands
- Argamon S, Koppel M, Fine J, Shimoni AR (2003) Gender, genre, and writing style in formal written texts. TEXT 23:321–346

- Asghari H, Mohtaj S, Fatemi O, Faili H, Rosso P, Potthast M (2016) Algorithms and corpora for persian plagiarism detection: overview of pan at fire 2016. In: Notebook papers of FIRE 2016, FIRE-2016, Kolkata, India, Dec 7–10, CEUR workshop proceedings, vol 1737, pp 135–144. CEUR-WS.org
- Bagnall D (2015) Author identification using multi-headed recurrent neural networks. In: Cappellato L, Ferro N, Gareth J, San Juan E (eds) Working notes papers of the CLEF 2015 evaluation labs
- Bagnall D (2016) Authorship clustering using multi-headed recurrent neural networks. In: Balog K, Cappellato L, Ferro N, Macdonald C (eds) CLEF 2016 evaluation labs and workshop – working notes papers. CEUR-WS.org
- Barrón-Cedeno A, Rosso P, Devi SL, Clough P, Stevenson M (2013) Pan@fire: overview of the cross-language Indian text re-use detection competition. In: Notebook papers of FIRE 2011, FIRE-2011, Mumbai, India, Dec 2–4
- Bensalem I, Boukhalfa I, Rosso P, Abouenour L, Darwish K, Chikhi S (2015) Overview of the AraPlagDet PAN@ FIRE2015 shared task on arabic plagiarism detection. In: Notebook papers of FIRE 2015, FIRE-2015, Gandhinagar, India, Dec 4–6, CEUR workshop proceedings, vol 1587, pp 111–122. CEUR-WS.org
- Burrows S, Potthast M, Stein B (2013) Paraphrase acquisition via crowdsourcing and machine learning. *Trans Intell Syst Technol (ACM TIST)* 4(3):43:1–43:21. <http://dx.doi.org/10.1145/2483669.2483676>
- ClueWeb09 (2009) The ClueWeb09 Dataset, 2009. <http://lemurproject.org/clueweb09/>
- Costa PT, McCrae RR (2008) The revised neo personality inventory (NEO-PI-R). The SAGE handbook of personality theory and assessment, vol 2. SAGE Publications, Los Angeles, pp 179–198
- Flores E, Rosso P, Moreno L, Villatoro-Tello E (2014) PAN@FIRE: overview of SOCO track on the detection of source code re-use. In: Notebook papers of FIRE 2014, FIRE-2014, Bangalore, India, Dec 5–7
- Flores E, Barrón-Cedeño A, Moreno L, Rosso P (2015) PAN@FIRE: overview of CL-SOCO track on the detection of cross-language source code re-use 1587:1–5
- Fréry J, Largeton C, Juganaru-Mathieu M (2014) UJM at CLEF in author identification. In: CLEF 2014 labs and workshops, notebook papers, CLEF and CEUR-WS.org
- Gollub T, Stein B, Burrows S (2012a) Ousting ivory tower research: towards a web framework for providing experiments as a service. In: Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval, ACM, pp 1125–1126
- Gollub T, Stein B, Burrows S, Hoppe D (2012b) Tira: Configuring, executing, and disseminating information retrieval experiments. In: Database and expert systems applications (DEXA), 2012 23rd international workshop on, IEEE, pp 151–155
- Gollub T, Potthast M, Beyer A, Busse M, Rangel F, Rosso P, Stamatatos E, Stein B (2013) Recent trends in digital text forensics and its evaluation: plagiarism detection, author identification, and author profiling. In: 4th international conference of CLEF on information access evaluation meets multilinguality, multimodality, and visualization, CLEF 2013, LNCS, vol 8138. Springer, New York, pp 53–58
- Gupta P, Clough P, Rosso P, Stevenson M (2012) Pan@fire: Overview of the cross-language Indian news story search (CL!NSS) track. In: Notebook papers of FIRE 2012, FIRE-2012, Kolkata, India, Dec 17–19
- Gupta P, Clough P, Rosso P, Stevenson M, Banchs RE (2013) Pan@fire: overview of the cross-language Indian news story search (CL!NSS) track. In: Notebook papers of FIRE 2013, FIRE-2013, Delhi, India, Dec 4–6
- Hagen M, Potthast M, Stein B (2015) Source retrieval for plagiarism detection from large web corpora: recent approaches. In: Working notes papers of the CLEF 2015 evaluation labs, CLEF and CEUR-WS.org, CEUR workshop proceedings. <http://www.clef-initiative.eu/publication/working-notes>
- Hagen M, Potthast M, Völske M, Gomoll J, Stein B (2016) How writers search: analyzing the search and writing logs of non-fictional essays. In: Kelly D, Capra R, Belkin N, Teevan

- J, Vakkari P (eds) Proceedings of the 1st ACM SIGIR conference on human information interaction and retrieval (CHIIR 16). ACM, New York, pp 193–202. <http://dx.doi.org/10.1145/2854946.2854969>
- Hagen M, Potthast M, Adineh P, Fatehifar E, Stein B (2017) Source retrieval for web-scale text reuse detection. In: Proceedings of the 26th ACM international conference on information and knowledge management (CIKM 17), ACM, New York
- Holmes J, Meyerhoff M (2003) The handbook of language and gender. Blackwell Handbooks in Linguistics. Wiley, Malden
- Inches G, Crestani F (2012) Overview of the international sexual predator identification competition at PAN-2012. In: Forner P, Karlgren J, Womser-Hacker C (eds) CLEF 2012 evaluation labs and workshop – working notes papers, 17–20 Sept, Rome, Italy
- Juola P, Stamatatos E (2013) Overview of the author identification task at PAN 2013. In: Working notes for CLEF 2013 conference
- Khonji M, Iraqi Y (2014) A slightly-modified GI-based author-verifier with lots of features (ASGALF). In: CLEF 2014 labs and workshops, notebook papers, CLEF and CEUR-WS.org
- Koppel M, Winter Y (2014) Determining if two documents are written by the same author. *J Am Soc Inf Sci Technol* 65(1):178–187
- Koppel M, Argamon S, Shimoni AR (2003) Automatically categorizing written texts by author gender. *Lit Ling Comput* 17(4): 401–412
- Koppel M, Schler J, Bonchek-Dokow E (2007) Measuring differentiability: unmasking pseudonymous authors. *J Mach Learn Res* 8:1261–1276
- López-Monroy AP, Montes-y Gómez M, Escalante HJ, Villaseñor-Pineda L, Stamatatos E (2015) Discriminative subprofile-specific representations for author profiling in social media. *Knowl-Based Syst* 89:134–147
- Maharjan S, Shrestha P, Solorio T, Hasan R (2014) A straightforward author profiling approach in MapReduce. In: Advances in artificial intelligence. Iberamia, pp 95–107
- Moreau E, Jayapal A, Lynch G, Vogel C (2015) Author Verification: Basic Stacked Generalization Applied To Predictions from a Set of Heterogeneous Learners. In: Cappellato L, Ferro N, Gareth J, San Juan E (eds) Working notes papers of the CLEF 2015 evaluation labs
- Pennebaker JW (2013) The secret life of pronouns: what our words say about us. Bloomsbury, New York
- Potthast M, Stein B, Eiselt A, Barrón-Cedeño A, Rosso P (2009) Overview of the 1st international competition on plagiarism detection. In: Stein B, Rosso P, Stamatatos E, Koppel M, Agirre E (eds) SEPLN 09 workshop on uncovering plagiarism, authorship, and social software misuse (PAN 09), CEUR-WS.org, pp 1–9. <http://ceur-ws.org/Vol-502>
- Potthast M, Barrón-Cedeño A, Eiselt A, Stein B, Rosso P (2010a) Overview of the 2nd international competition on plagiarism detection. In: Braschler M, Harman D, Pianta E (eds) Working notes papers of the CLEF 2010 evaluation labs. <http://www.clef-initiative.eu/publication/working-notes>
- Potthast M, Stein B, Barrón-Cedeño A, Rosso P (2010b) An evaluation framework for plagiarism detection. In: Huang CR, Jurafsky D (eds) 23rd international conference on computational linguistics (COLING 10). Association for computational linguistics, Stroudsburg, Pennsylvania, pp 997–1005
- Potthast M, Eiselt A, Barrón-Cedeño A, Stein B, Rosso P (2011) Overview of the 3rd international competition on plagiarism detection. In: Petras V, Forner P, Clough P (eds) Working notes papers of the CLEF 2011 evaluation labs. <http://www.clef-initiative.eu/publication/working-notes>
- Potthast M, Gollub T, Hagen M, Graßegger J, Kiesel J, Michel M, Oberländer A, Tippmann M, Barrón-Cedeño A, Gupta P, Rosso P, Stein B (2012a) Overview of the 4th international competition on plagiarism detection. In: Forner P, Karlgren J, Womser-Hacker C (eds) Working notes papers of the CLEF 2012 evaluation labs. <http://www.clef-initiative.eu/publication/working-notes>
- Potthast M, Hagen M, Stein B, Graßegger J, Michel M, Tippmann M, Welsch C (2012b) ChatNoir: a search engine for the ClueWeb09 corpus. In: Hersh B, Callan J, Maarek Y, Sanderson M



- (eds) 35th international ACM conference on research and development in information retrieval (SIGIR 12), ACM, p 1004. <http://dx.doi.org/10.1145/2348283.2348429>
- Potthast M, Gollub T, Hagen M, Tippmann M, Kiesel J, Rosso P, Stamatatos E, Stein B (2013a) Overview of the 5th international competition on plagiarism detection. In: Forner P, Navigli R, Tufis D (eds) Working notes papers of the CLEF 2013 evaluation labs. <http://www.clef-initiative.eu/publication/working-notes>
- Potthast M, Hagen M, Völske M, Stein B (2013b) Crowdsourcing interaction logs to understand text reuse from the web. In: Fung P, Poesio M (eds) Proceedings of the 51st annual meeting of the association for computational linguistics (ACL 13). Association for computational linguistics, pp 1212–1221. <http://www.aclweb.org/anthology/P13-1119>
- Potthast M, Gollub T, Rangel F, Rosso P, Stamatatos E, Stein B (2014a) Improving the reproducibility of pan's shared tasks: Plagiarism detection, author identification, and author profiling. In: 5th international conference of CLEF on information access evaluation meets multilinguality, multimodality, and interaction, CLEF 2014. LNCS, vol 8685. Springer, New York, pp 268–299
- Potthast M, Hagen M, Beyer A, Busse M, Tippmann M, Rosso P, Stein B (2014b) Overview of the 6th international competition on plagiarism detection. In: Cappellato L, Ferro N, Halvey M, Kraaij W (eds) Working notes papers of the CLEF 2014 evaluation labs, CLEF and CEUR-WS.org, CEUR workshop proceedings. <http://www.clef-initiative.eu/publication/working-notes>
- Potthast M, Göring S, Rosso P, Stein B (2015) Towards data submissions for shared tasks: first experiences for the task of text alignment. In: Working notes papers of the CLEF 2015 evaluation labs, CLEF and CEUR-WS.org, CEUR workshop proceedings. <http://www.clef-initiative.eu/publication/working-notes>
- Potthast M, Rangel F, Tschuggnall M, Stamatatos E, Rosso P, Stein B (2017) Overview of PAN'17: author identification, author profiling, and author obfuscation. In: 8th international conference of CLEF on experimental IR meets multilinguality, multimodality, and visualization, CLEF 2017, LNCS, vol 10456. Springer, New York, pp 275–290
- Rammstedt B, John O (2007) Measuring personality in one minute or less: A 10 item short version of the big five inventory in English and German. *J Res Pers* 203–212
- Rangel F, Rosso P (2015) On the multilingual and genre robustness of emographs for author profiling in social media. In: 6th international conference of CLEF on experimental IR meets multilinguality, multimodality, and interaction, LNCS, vol 9283. Springer, New York, pp 274–280
- Rangel F, Rosso P (2016) On the impact of emotions on author profiling. *Inf Process Manage* 52(1):73–92
- Rangel F, Rosso P, Moshe Koppel M, Stamatatos E, Inches G (2013) Overview of the author profiling task at pan 2013. In: Forner P, Navigli R, Tufis D (eds) CLEF 2013 labs and workshops, notebook papers, vol 1179. CEUR-WS.org
- Rangel F, Rosso P, Chugur I, Potthast M, Trenkmann M, Stein B, Verhoeven B, Daelemans W (2014) Overview of the 2nd author profiling task at PAN 2014. In: Cappellato L, Ferro N, Halvey M, Kraaij W (eds) CLEF 2014 labs and workshops, notebook papers, vol 1180. CEUR-WS.org
- Rangel F, Rosso P, Potthast M, Stein B, Daelemans W (2015) Overview of the 3rd author profiling task at pan 2015. In: Cappellato L, Ferro N, Jones G, San Juan E (eds) CLEF 2015 labs and workshops, notebook papers. CEUR workshop proceedings, vol 1391. CEUR-WS.org
- Rangel F, González F, Restrepo F, Montes M, Rosso P (2016a) Pan at fire: Overview of the PR-SOCO track on personality recognition in source code. Notebook papers of FIRE 2016, FIRE-2016, Kolkata, India, Dec 7–10, CEUR workshop proceedings, vol 1737, pp 1–5. CEUR-WS.org
- Rangel F, Rosso P, Verhoeven B, Daelemans W, Potthast M, Stein B (2016b) Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In: Working notes papers of the CLEF 2016 Evaluation labs, CLEF and CEUR-WS.org, CEUR workshop proceedings

- Rangel F, Rosso P, Potthast M, Stein B (2017) Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in twitter. Working notes papers of the CLEF
- Rosso P, Rangel F, Potthast M, Stamatatos E, Tschuggnall M, Stein B (2016) Overview of the PAN'2016 - new challenges for authorship analysis: Cross-genre profiling, clustering, diarization, and obfuscation. In: 7th international conference of CLEF on Experimental IR meets multilinguality, multimodality, and interaction, CLEF 2016, LNCS, vol 9822. Springer, New York, pp 332–350
- Sadat F, Kazemi F, Farzindar A (2014) Automatic identification of arabic language varieties and dialects in social media. In: Proceedings of SocialNLP, p 22
- Schler J, Koppel M, Argamon S, Pennebaker JW (2006) Effects of age and gender on blogging. In: AAAI spring symposium: computational approaches to analyzing weblogs, AAAI, pp 199–205
- Seidman S (2013) Authorship verification using the impostors method. In: Forner P, Navigli R, Tufis D (eds) CLEF 2013 Evaluation labs and workshop – Working notes papers
- Stamatatos E (2009) A survey of modern authorship attribution methods. *J Am Soc Inf Sci Technol* 60:538–556
- Stamatatos E (2011) Plagiarism detection using stopword n-grams. *J Am Soc Inf Sci Technol* 62(12):2512–2527. <http://dx.doi.org/10.1002/asi.21630>
- Stamatatos E, Daelemans W, Verhoeven B, Stein B, Potthast M, Juola P, Sánchez-Pérez MA, Barrón-Cedeño A (2014) Overview of the author identification task at PAN 2014. In: Working notes for CLEF 2014 conference, pp 877–897
- Stamatatos E, Daelemans W, Verhoeven B, Juola P, López-López A, Potthast M, Stein B (2015a) Overview of the author identification task at PAN 2015. In: Working notes of CLEF 2015 - conference and labs of the evaluation forum
- Stamatatos E, Potthast M, Rangel F, Rosso P, Stein B (2015b) Overview of the pan/clef 2015 evaluation lab. In: 6th international conference of CLEF on experimental IR meets multilinguality, multimodality, and interaction, CLEF 2015. LNCS, vol 9283. Springer, New York, pp 518–538
- Stamatatos E, Tschuggnall M, Verhoeven B, Daelemans W, Specht G, Stein B, Potthast M (2016) Clustering by authorship within and across documents. In: Working notes papers of the CLEF 2016 Evaluation labs, CLEF and CEUR-WS.org, CEUR workshop proceedings, vol 1609. <http://ceur-ws.org/Vol-1609/>
- Stein B, Meyer zu Eißben S, Potthast M (2007) Strategies for retrieving plagiarized documents. In: Clarke C, Fuhr N, Kando N, Kraaij W, de Vries A (eds) 30th International ACM conference on research and development in information retrieval (SIGIR 07). ACM, New York, pp 825–826. <http://dx.doi.org/10.1145/1277741.1277928>
- Stein B, Lipka N, Prettenhofer P (2011) Intrinsic plagiarism analysis. *Lang Resour Eval (LRE)* 45(1):63–82. <http://dx.doi.org/10.1007/s10579-010-9115-y>
- Tschuggnall M, Stamatatos E, Verhoeven B, Daelemans W, Specht G, Stein B, Potthast M (2017) Overview of the author identification task at PAN-2017: style breach detection and author clustering. In: Working notes papers of the CLEF 2017 evaluation labs, CLEF and CEUR-WS.org. CEUR workshop proceedings
- Weren E, Kauer A, Mizusaki L, Moreira V, de Oliveira P, Wives L (2014) Examining multiple features for author profiling. *J Inf Data Manage* 5:266–279

# RepLab: An Evaluation Campaign for Online Monitoring Systems



Jorge Carrillo-de-Albornoz, Julio Gonzalo, and Enrique Amigó

**Abstract** Over a period of 3 years, RepLab was a CLEF initiative where computer scientists and online reputation experts worked together to identify and formalize the computational challenges in the area of online reputation monitoring. Two main results emerged from RepLab: a community of researchers engaged in the problem, and an extensive Twitter test collection comprising more than half a million expert annotations, which cover many relevant tasks in the field of online reputation: named entity resolution, topic detection and tracking, reputational alerts identification, reputational polarity, author profiling, opinion makers identification and reputational dimension classification. It has probably been one of the CLEF labs with a larger set of expert annotations provided to participants in a single year, and one of the labs where the target user community has been more actively engaged in the evaluation campaign. Here we summarize the design and results of the Replab campaigns, and also report on research that has built on RepLab datasets after completion of the 3-year competition cycle.

## 1 Introduction

Corporate reputation has been an intense subject of study in the last 30 years. It has been shown to be one of the most valuable assets of companies and organizations (Doorley and Garcia 2011). Research confirms its great influence on the behavior of all the stakeholders. To begin with, companies with better reputations engender loyalty in consumers across several generations and countries (Alsop 2006). Second, a solid reputation adds value to the actual worth of a company and awakens the interest of investors (Kreps and Wilson 1982). Finally, having a good reputation is crucial to attract highly qualified employees and thereby become more efficient and productive (Chong and Tan 2010). It is only logical that companies and

---

J. Carrillo-de-Albornoz · J. Gonzalo (✉) · E. Amigó  
NLP & IR Group at UNED, Madrid, Spain  
e-mail: [jcalbornoz@lsi.uned.es](mailto:jcalbornoz@lsi.uned.es); [julio@lsi.uned.es](mailto:julio@lsi.uned.es); [enrique@lsi.uned.es](mailto:enrique@lsi.uned.es)

organizations dedicate considerable resources to the management of such a key component of their business development.

Reputation management involves activities that aim at building and preserving a company's reputation. In the past, it was predominantly static, and mainly comprised building an attractive image via marketing campaigns and carefully planned corporate messages. Nowadays, social media have radically changed the traditional reputation management model, giving rise to new channels of communication between companies and their audience. Current technology applications provide users with a wide access to information, enabling them to share it instantly and 24 h a day due to constant connectivity. Information, including users' opinions about people, companies or products, is quickly spread over large communities. In this setting, every move of a company and every act of a public figure, are subject, at all times, to the scrutiny of a powerful global audience. The control of information about public figures and organizations has at least partly moved from them to users and consumers (Hoffman 2008; Jansen et al. 2009b; Glance et al. 2005). So that, for an effective Online Reputation Management (ORM), this constant flow of online opinions needs to be watched.

While traditional reputation analysis is mostly manual, online media make it possible to process, understand and aggregate large streams of facts and opinions about companies and individuals in an automatic manner. In this context, Natural Language Processing plays a key, enabling role and we are already witnessing an unprecedented demand for text mining software for ORM. Although opinion mining has made significant advances in the last few years, most work has been focused on products. However, mining and interpreting opinions about companies and individuals is, in general, a much harder and less understood problem since, unlike products or services, opinions about people and organizations cannot be structured around any fixed set of features or aspects, requiring a more complex modelling of these entities.

RepLab was an initiative promoted by the EU project LiMoSINE,<sup>1</sup> and aimed at structuring research on reputation management as a series of evaluation campaigns in which task design and evaluation methodologies are jointly developed by researchers and the target user communities (reputation management experts). The focus was on detecting challenges and opportunities for language technologies in online reputation monitoring problems, to define appropriate evaluation methodologies, build evaluation test collections with reference annotations provided by reputation experts, and run shared tasks on these collections with research labs from academia and industry.

Replab focused on Twitter data, and was designed to run in a 3-year cycle. The first evaluation campaign was held as a CLEF 2012 activity, and focused on a pilot task around the daily work of reputation experts. The monitoring task for analysts, as studied in RepLab, essentially consisted of searching the stream of tweets for potential mentions to the entity, filtering those that do refer to the entity,

---

<sup>1</sup>[http://cordis.europa.eu/fp7/ict/language-technologies/project-limosine\\_en.html](http://cordis.europa.eu/fp7/ict/language-technologies/project-limosine_en.html).

detecting topics (i.e., clustering tweets by subject) and ranking them based on the degree to which they are potential reputation alerts (i.e., issues that may have a substantial positive or negative impact on the reputation of the entity, and must be handled by reputation management experts). RepLab 2013 kept the same tasks and worked on producing a much larger, expert annotated dataset which comprises more than half a million manual annotations on tweets related to companies, universities and music bands. Finally, RepLab 2014 focused on two additional aspects of reputation analysis (reputation dimensions classification and author profiling) that complemented the tasks tackled in the previous campaigns.

In this chapter, we summarize the organization and results of RepLab evaluation campaigns, explore how RepLab datasets have been used to advance the state of the art from the end of RepLab up to now (2019), and discuss the lessons learnt along the way.

The chapter is organized as follows: Sect. 2 summarizes the three evaluation campaigns, including the participants, datasets and evaluation methodologies. Section 3 describes the tasks and their outcome. Section 4 summarizes post-RepLab research. Finally, Sect. 5 discusses the main lessons learned.

## 2 RepLab Evaluation Campaigns

RepLab was a competitive evaluation exercise supported by the EU project LiMo-SiNE. It aimed at encouraging research on Online Reputation Management and providing a framework for collaboration between academia and practitioners. A crucial feature of RepLab was that task design was jointly carried out by researchers and the target user community (reputation management experts). All evaluation campaigns were co-organized by three members of the Limosine project: Universidad Nacional de Educacion a Distancia (UNED) and University of Amsterdam (UvA) as academic partners, and the reputational experts of the consultancy firm Llorente & Cuenca and Yahoo! Research as industrial partners. The RepLab evaluation campaigns were carried out during years 2012, 2013 and 2014.

### 2.1 *Problem Setup: Tasks and Metrics*

The working scenario for RepLab is that of reputation experts constantly tracking and annotating information about a client (an entity that can be an organization, brand, individual, etc.). We focused on Twitter data for two reasons: it is a primary source to be tracked by online reputation experts, as it tends to be the online place where things happen first; it has a more open nature than other social networks (such as facebook), and therefore there are less privacy issues when downloading and working with Twitter data. Although it would have been great to work on several

social media, it proved too complex for the scope of our 3-year evaluation cycle and the resources available.

In the basic workflow of an online reputation expert working for a client, RepLab organizers identified several relevant subtasks where automation could substantially speed up the process: finding out whether tweets containing the entity name were actually about the entity (*filtering* or *disambiguation* task), annotating their reputation polarity (does the content have negative or positive implications for the reputation of the entity?), finding out which are the topics discussed about the entity, which of these topics are reputation alerts, what are the reputational dimensions of the entity involved in a topic, identifying whether tweet authors were influencers in the activity domain of the entity, etc.

Each task corresponds to a particular abstract problem, as for example binary classification (*filtering*), three-level classification (polarity and priority), clustering (topic detection) or ranking (author influence). A common feature of the data for all tasks is that the classes, levels or clusters tend to be unbalanced. This entails challenges both for the systems and for the definition of the evaluation methodology. First, in classification tasks, a non informative system (i.e., all tweets to the same class) can achieve high scores without providing useful information. Second, in multi-class classification tasks, a system could sort tweets correctly without a perfect correspondence between predicted and true tags. Third, an unbalanced cluster distribution across entities produces an important trade-off between precision/recall oriented evaluation metrics (precision or cluster entropy versus recall or class entropy) and that makes the measure combination function crucial for system ranking.

We also wanted to have a measure of the quality of a reputation monitoring system as a whole, i.e. as a result of the combination of all the above individual tasks. We focused on our so-called “full monitoring task” as a combination of filtering (classify relatedness content), clustering (into topically-related texts) and ranking (clusters must be ranked by priority). To our knowledge, there was no standard evaluation measure for this type of combined problem. We dedicated part of our efforts to design a suitable evaluation measure for this problem. We started by defining a general “document organization problem” that subsumes clustering, retrieval and filtering. We defined an evaluation measure for this combined problem that satisfies all desirable properties for each of the subsumed tasks (expressed as formal constraints). This measure is the combination (via a weighted harmonic mean) of *Reliability* and *Sensitivity* (Amigó et al. 2013), defined as Precision and Recall of the binary document relationships predicted by a system on the set of relationships established in the gold standard, with a specific weighting scheme.

In evaluation, there is usually a trade-off between interpretability and strictness. For instance, Accuracy is easy to interpret: it simply reports how frequently the system makes the correct decision. However, it is of little use with unbalanced test sets. For instance, returning all tweets in the same class, cluster or level, may have high accuracy if the set is unbalanced. Other measures based on information theory are stricter when penalizing non informative outputs, but at the cost of interpretability. In the RepLab evaluation campaigns we employed Accuracy as a

highly interpretable measure, and the combination of Reliability and Sensitivity (R&S) as a strict, theoretically sound measure.

R and S are combined with the F measure, i.e. a weighted harmonic mean of R and S. This combining function is grounded on measurement theory, and satisfies a set of desirable constraints. One of the most useful is that a low score according to any individual measure penalizes the combined score. However, specially in clustering tasks, the F measure is seriously affected by the relative weight of partial measures (the  $\alpha$  parameter). In order to solve this we complement the evaluation results with the Unanimous Improvement Ratio, which has been proved to be the only weighting independent combining criterion (Amigó et al. 2011). UIR is computed over the test cases (entities in RepLab) in which all measures corroborate a difference between runs. Being  $S_1$  and  $S_2$  two runs and  $N_{>\forall}(S_1, S_2)$  the amount of test cases for which  $S_1$  improves  $S_2$  for all measures:

$$UIR(S_1, S_2) = \frac{N_{>\forall}(S_1, S_2) - N_{>\forall}(S_2, S_1)}{\text{Amount of cases}}$$

Finally, we also dealt with the problem of identifying influencers in a given activity domain. This can be modeled as a binary classification task (each Twitter author must be categorized as influencer or non influencer) or as a ranking task (the system must return a list of authors with decreasing probability of being influencers). The main difference with a standard retrieval task is that the ratio of relevant authors turned out to be higher than the typical ratio of relevant documents in IR. Another differentiating characteristic is that the set of potentially influential authors is rather small, while information retrieval data sets usually consist of millions of documents. This has implications for the evaluation methodology. Most Information Retrieval measures reflect the fact that users are less likely to explore items which are deeper in the results list. It is not trivial to estimate how deep in the ranking reputation experts are expected to go; but it is obviously deeper than in a typical search, as their goal is to find as many opinion makers as possible. Hence, we decided to use *MAP* (*Mean Average Precision*), which is recall oriented and also considers the relevance of authors at lower ranks.

## 2.2 RepLab Datasets

RepLab comprises three different datasets built in the three evaluation campaigns (2012, 2013 and 2014):

- RepLab 2012 focused on the scenario of an online application where the user types in an entity name, and the system retrieves and organizes textual information about the entity. In this scenario, it cannot be assumed that there is entity-specific training material for the system. Therefore, training and test sets refer to different entities, and systems must be able to properly generalize on the

training data. Tweets in English and Spanish, containing the name of an entity of interest, were annotated according to several subtasks: whether the tweet talks about the entity or not, what is the reputational polarity of the tweet, which are the tweets talking about the same issue, and what is the relative importance of each issue from a reputational perspective.

- RepLab 2013 focused on the scenario where systems must help online reputation experts, who are constantly tracking and annotating information about a client (an organization, brand, individual, etc.). In this case, it is reasonable to assume that systems have previously annotated material about each entity. Tasks were the same as in 2012, and the main difference in design with respect to the 2012 dataset is that in this case, training and test materials refer to the same set of entities.
- RepLab 2014 used the same set of tweets as in 2013, expanding the annotations to two additional tasks: author profiling (who are the opinion makers and what type of activity do they have) and dimension categorization (what reputational dimension of the entity is affected by a tweet?).

RepLab datasets focus on Twitter data in English and Spanish. The balance between both languages depends on the availability of data for each of the entities included in the dataset. The main reason for choosing Twitter is that it currently constitutes the first source for the latest news (Krishnamurthy et al. 2008), due to its ubiquitous and real-time nature, and had been little studied for automating the ORM process (Li and Li 2013; Jansen et al. 2009a).

The **RepLab 2012** manual annotations were provided by online reputation management experts from the Public Relations consultancy Llorente & Cuenca. Such annotations are much more costly than a crowdsourcing alternative, but they have the crucial advantage that data serves not only to evaluate systems, but also to understand the concept of reputation from the perspective of professional practitioners. The RepLab 2012 training dataset consists of at least 30,000 tweets crawled per each company name, for six companies<sup>2</sup> using the company name as query, in English and Spanish. The time span and the proportion between English and Spanish tweets depends on the company. For each company's timeline, 300 tweets (approximately in the middle of the timeline) were manually annotated by reputation management experts. This is the *labelled* dataset. The rest (around 15,000 unannotated tweets before and after the annotated set, for each company), is the *background* dataset. Tweets in the background set have not been annotated.

Test data are identical to training data, for a different set of 31 companies.<sup>3</sup> The tweets were crawled using the company identifier as query. There are between 19,400 and 50,000 tweets per company name, in English and Spanish. Similarly

---

<sup>2</sup>Training set: Apple, Lufthansa, Alcatel, Armani, Marriott, Barclays.

<sup>3</sup>Test set: Telefonica, BBVA, Repsol, Indra, Endesa, BME, Bankia, Iberdrola, "Banco Santander", Mediaset, IAG, Inditex, Mapfre, Caixabank, "Gas Natural", Yahoo, Bing, Google, ING, "Bank of America", Blackberry, BMW, BP, Chevrolet, Ferrari, Fiat, VW, Wilkinson, Gillette, Nivea, Microsoft.



to the training set, the time span, and the proportion between English and Spanish tweets here depends on the company. For each company's timeline, approximately in the middle, between 190 and 400 tweets are annotated by reputation management experts. The actual size for each entity depends on the availability of tweets at evaluation time for each company. "Labelled" tweets will be used to evaluate systems. Again, for each company the "background" dataset contains the tweets before and after the annotated test set.

The labelled data is annotated as follows by the ORM experts:

- Each tweet is first annotated with relatedness information (*yes*, if the tweet refers to the entity analysed, *no* otherwise).
- Those tweets related with the company are then labelled according to its polarity for reputation (does the tweet content have *positive/neutral/negative* implications for the company's reputation?).
- Tweets are clustered topically (using topic labels).
- Clusters are annotated for priority (does the cluster topic demand urgent attention from the point of view of reputation management?), in three levels (reputation alert, mildly important, unimportant).

Note that: (1) unlike many test collections, in RepLab 2012 the test set is significantly larger than the trial set, which is too small to be used as proper training corpora; (2) companies in the trial and test collections are different; therefore, systems cannot individually learn features for each company; they must learn features at a higher level of generalization. Both design decisions were intended to avoid a large set of systems that blindly apply Machine Learning machinery, and to push participants into creative solutions to the problem.

In its second year, **RepLab 2013** focused on the daily tasks of an online reputation management expert. The collection comprises tweets mentioning 61 different entities from four domains: automotive, banking, universities and music. The domain selection was intended to offer a variety of scenarios for reputation studies. To this aim, we included (1) entities whose reputation largely relies on their products (automotive), (2) entities for which transparency and ethical side of their activity are the most decisive reputation factors (banking); (3) entities for which their reputation depends on a very broad and intangible set of products (universities) and, finally, (4) entities for which their reputation depends almost equally on their products and personal qualities (music bands and artists).

Crawling was performed from 1 June, 2012 up to 31 Dec, 2012, using each entity's canonical name as query. For each entity, at least 2200 tweets were collected: the first 700 were reserved for the training set and the last 1500 for the test collection. This distribution was set in this way to obtain a temporal separation (of several months) between the training and test data. The corpus also comprises additional background tweets for each entity (up to 50,000, with a large variability across entities). These are the remaining tweets situated between the training (earlier tweets) and test material (the latest tweets) in the timeline. These data sets were manually labelled by thirteen annotators who were trained, guided and

constantly monitored by experts from Llorente & Cuenca. Each tweet is annotated as follows:

- RELATED/UNRELATED: the tweet is/is not about the entity.
- POSITIVE/NEUTRAL/NEGATIVE: the information contained in the tweet has positive, neutral or negative implications for the entity's reputation.
- Identifier of the topic cluster the tweet has been assigned to.
- ALERT/MILDLY IMPORTANT/UNIMPORTANT: the priority of the topic cluster the tweet belongs to.

The RepLab 2013 dataset is the largest of the three produced for the RepLab campaigns, and consists of more than 142,000 labelled tweets in English and Spanish, containing more than 500,000 manual labels overall. The total annotation workload was of 21 person-month. The dataset is divided in 45,679 tweets for the training set, and 96,848 tweets for the test set.

Finally, **RepLab 2014** comprises two different datasets: the *Reputation Dimensions* Dataset and the *Author Profiling* Dataset. The first one provides additional annotations to the RepLab 2013 tweet dataset, with over 48,000 manually labelled English and Spanish tweets related to 31 entities from the automotive and banking domains. The training set is composed of 15,562 Twitter posts and 32,446 tweets are reserved for the test set. Both data sets were manually labelled by annotators trained and supervised by experts in ORM from the online division of Llorente & Cuenca.

The tweets were classified according to the RepTrak dimensions<sup>4</sup>: *Performance, Product and Services, Leadership, Citizenship, Governance, Workplace, and Innovation*. In case a tweet cannot be categorised into any of these dimensions, it was labelled as "Undefined". As in the RepLab 2013 dataset, the reputation dimensions corpus also comprises additional background tweets for each entity (up to 50,000, with a large variability across entities). These are the remaining tweets temporally situated between the training (earlier tweets) and test material (the latest tweets) in the timeline.

The Author Profiling data collection contains over 7000 Twitter profiles (all with at least 1000 followers) from the automotive and banking domains, together with an additional set of miscellaneous profiles (the idea of this extra set is to evaluate if approaches designed for a specific domain are suitable for a broader multi-domain scenario). Each profile contains (1) its screen name; (2) its profile URL, and (3) the most recent 600 tweets published by the author at crawling time.

The collection was split into training and test sets: 2500 profiles in the training set and 4991 profiles in the test set. Reputation experts from Llorente & Cuenca provided manual annotations for two subtasks: Author Categorisation and opinion makers identification. For the first task, author profiles are categorized according to the following options: *company* (i.e., corporate accounts of companies), *professional, celebrity, employee, stockholder, journalist, investor, sportsman, public*

---

<sup>4</sup><https://www.reputationinstitute.com/about-reputation-institute/the-retrak-framework>.

*institution*, and *non-governmental organisation (NGO)*. For the second task, reputation experts manually identified opinion makers (i.e., authors with reputational influence) and annotated them as “Influencer”. The profiles that were not considered opinion makers were labelled as “Non-Influencer”. Profiles that could not be clearly classified into one of these categories were labelled as “Undecidable”.

Note that the current amount of available tweets may be lower, as some posts may have been deleted or made private by the authors: in order to respect Twitter’s terms of service, we did not provide the contents of the tweets, but only tweet ids and screen names.

### 2.3 Participation

Overall, the RepLab evaluation campaigns attracted a remarkable number of research teams. A total of 132 groups registered for one or more tasks (39 in 2012, 44 in 2013 and 49 in 2014). Out of them, 42 groups (from 15 countries) were able to submit runs. Broadly speaking, the main focus of interest was the filtering task, which attracted a total of 23 participants (9 in 2012 and 14 in 2013), followed by the polarity for reputation task, with 21 teams submitting runs (10 in 2012 and 11 in 2013).

The topic detection and topic priority tasks attracted less participation, with eight teams submitting runs (3 in 2012 and 5 in 2013). In 2014, eight groups participated in the Reputation Dimensions task and five groups submitted their results to the Author Profiling challenge (all of them attempted the opinion maker identification subtask, and all but one the author categorization subtask).

## 3 Tasks and Results

Typically, an online reputation analyst periodically performs the following tasks (with the assistance of more or less sophisticated software):

- Starts with a set of queries that cover all possible ways of referring to the client.
- Takes the set of results and filters out irrelevant content.
- Identifies the different issues (topics) in relation with the client, and groups tweets accordingly.
- Evaluates the reputational priority of each issue, establishing at least three categories: reputation alerts (which demand immediate attention), relevant topics (that the company must be aware of), and unimportant content (refers to the entity, but does not have consequences from a reputational point of view).
- Produces a reputation report for the client, summarizing the results of the analysis.

Figure 1 describes the main steps carried out during the annotation process for reputation monitoring. The process starts by selecting one of the entities assigned to the expert. In the system, each entity has a list of tweets that the expert has to annotate manually. The expert processes tweets sequentially: first, she decides whether the tweet does refer to the entity of interest or not. If the tweet is unrelated to the entity, the annotation process for the tweet finishes and the expert continues with the next tweet in the list. Otherwise, the polarity and topic annotations follow. Polarity annotation consists in deciding whether the tweet may affect positively or negatively the reputation of the entity.

Topic annotation consists of identifying the aspects and events related to the entity that the tweet refers to. If the tweet refers to an already identified topic, the tweet is assigned to it. Otherwise, the expert defines a new topic. A topic receives a label that summarizes what the topic is about, and it is also classified in a priority scale (Alert, Medium or Low). When the tweet is assigned to a topic, the annotation of the current tweet is finished.

In this process, reputational experts take into account several aspects of the tweet in order to determine the different labels described above. Some of them include the novelty of the topic (already known issues tend to be less relevant), centrality (whether the company is the main focus of the content), its potential impact, the company dimensions affected by the text, and the profile of the author (her influence and her role). The first three features focus on the tweet itself, and aim to better understand it as a whole. On the other hand, the reputation dimensions contribute to a better understanding of the topic of a tweet or group of tweets, whilst author profiling provides important information for priority ranking of tweets, as certain characteristics of the author can make a tweet (or a group of tweets) an alert, requiring special attention of reputation experts. The types of opinion holders and the company dimensions are standard annotations (RepTrack guidelines<sup>5</sup>), while the influence of the author must be interpreted by the expert for each specific domain.

The next subsections describe the different text understanding tasks that are involved in this labelling process.

### ***3.1 Named Entity Disambiguation***

Reputation monitoring is strongly recall-oriented (nothing relevant to the company should be missed), and therefore queries are usually short and ambiguous, and may generate a lot of noise (consider Blackberry, Orange and Apple, just to mention a few companies whose names are also words for fruits). An automatic solution to this initial filtering problem would already have a major impact on the budget needed to monitor online information. An evaluation campaign focused on company name disambiguation in Twitter (WePS-3) already proved that this is not a trivial problem:

---

<sup>5</sup><https://www.reputationinstitute.com/about-reputation-institute/the-retrak-framework>.

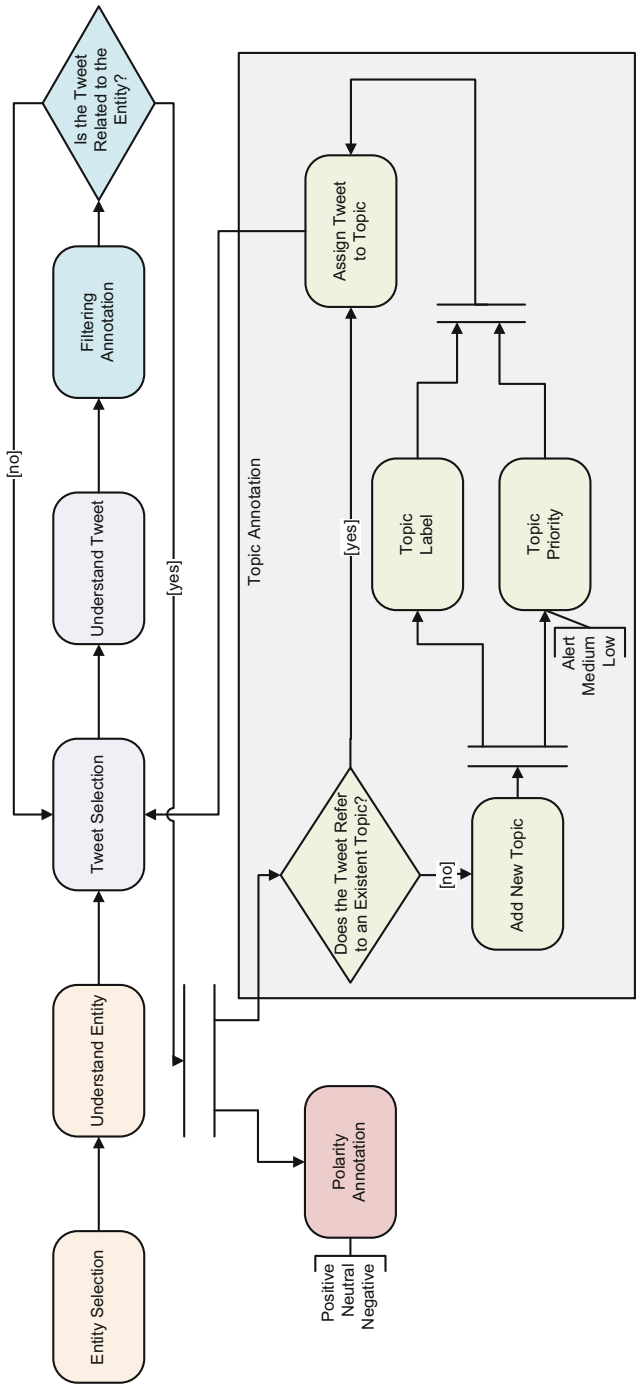


Fig. 1 Workflow of the online reputation monitoring annotation process

the best fully automatic system had a performance of 75% accuracy, which is not impressive considering that a random baseline gets 50%.

Systems were asked to determine which tweets are related to the entity and which are not. For instance, distinguishing between tweets that contain the word “Stanford” referring to the University of Stanford and filtering out tweets about Stanford as a place. Manual annotations were provided with two possible values: related/unrelated. As explained above, Reliability and Sensitivity were used for this task; for a filtering task, they correspond to the products of precision in both classes and the product of recall scores, respectively. Systems were ranked by the harmonic mean of their Reliability and Sensitivity ( $F(R,S)$ ), and Accuracy was also reported, although classes are imbalanced to different degrees depending on the company.

Looking at the top performing systems for RepLab 2012 in terms of  $F(R, S)$  (0,26) (Villena-Román et al. 2012) and accuracy (0,81 for a baseline of 0,71) (Kaptein 2012), it seems that there is still a wide margin to improve system performance. Note that the Replab setting in this first edition was, however, the most challenging setting for filtering algorithms, because the training set is small and does not use the same set of entities as the test set. In the RepLab 2013 edition, training and test sets referred to the same company, which led to better system performance. Best systems achieved  $F(R,S)$  of 0,49 (Filgueiras and Amir 2013) and accuracy of 0,93 (for a baseline of 0,87) (Hangya and Farkas 2013), making filtering as a real candidate for a fully automatic task.

### 3.2 *Polarity for Reputation*

Does the information (facts, opinions) in the text have positive, negative, or neutral implications for the image of the company? This problem is related to sentiment analysis and opinion mining, but has substantial differences. First, when analyzing polarity for reputation, both facts and opinions may have reputational polarity. For instance, “Barclays plans additional job cuts in the next 2 years” is a fact with negative implications for reputation. Therefore, systems were not explicitly asked to classify tweets as factual vs. opinionated: the goal was to find polarity for reputation, that is, what implications a piece of information might have on the reputation of a given entity, regardless of whether the content is opinionated or not. Second, negative sentiments do not always imply negative polarity for reputation and vice versa. For instance, “R.I.P. Michael Jackson. We’ll miss you” has a negative associated sentiment (sadness, deep sorrow), but a positive implication for the reputation of Michael Jackson. And the other way around, a tweet such as “I LIKE IT.... NEXT... MITT ROMNEY... Man sentenced for hiding millions in Swiss bank account” has a positive sentiment (joy about a sentence) but has a negative implication for the reputation of Mitt Romney.

While only a small percentage (around 15%) of generic tweets have sentiment polarity, tweets talking about companies and celebrities are highly polar from the point of view of their reputational implications. According to the reputational

experts, tweets in our collections have positive or negative polarity in 67% of the cases in the 2012 RepLab collection and 73% in the RepLab 2013 collection.

Regarding the results, again, the task was much more challenging in 2012, with the best systems achieving 0,40 F(R,S) (Villena-Román et al. 2012) and 0,49 accuracy (Carrillo-de-Albornoz et al. 2012), respectively. According to F (R, S), detecting polarity seems to be—surprisingly—less challenging than the filtering task (0,48 is the top result for polarity and 0,26 the top result for filtering). Note that accuracy tells a very different story, because it rewards baseline “all positive” in the filtering task, while for the polarity task, as it has three relatively balanced classes, gives lower results for the baselines. In the 2013 scenario, the results of the best participants (Hangya and Farkas 2013) considerably outperform the best 2012 results in terms of accuracy (0,69), but not in terms of F(R,S) (0,38). This probably indicates that in 2013 systems were learning about the majority class, but were not generalizing adequately.

### 3.3 *Topic Detection*

The ability of distinguishing the different issues people are talking about, grouping together texts that refer to the same issue, tracking issues along time, detecting novel topics, etc., is crucial for automatic reputation management and also for assisting reputation experts and facilitating their analysis tasks.

Systems are asked to cluster related tweets about the entity by topics, with the goal of identifying subjects/events/conversations and their relative size. Topic detection is, therefore, a clustering task that was evaluated according to R&S, which for the clustering problem corresponds to Bcubed precision and Recall (Amigó et al. 2009).

In terms of clustering, the three participant groups in 2012 (Martin et al. 2012; Qureshi et al. 2012; Balahur and Tanev 2012) achieved a similar performance (F(R,S) between 0,38 and 0,40), below the baseline algorithm provided by the organizers (Hierarchical Agglomerative Clustering) with thresholds 0, 10, 20. This was an indication that systems were not yet substantially contributing to solve the problem. Note that the topics are of a rather small size when compared to other clustering problems, and standard methods that require more data, such as LDA, turned out not to be effective in this context. Of course this difference has to be put in perspective: we have implemented the baseline for eleven different values of the stopping threshold, which means that the best performing baseline had an “oracle” effect, i.e., it is using the optimal threshold setting for the test corpus. The best results in 2013 (0,33 and 0,29 F(R,S), achieved by Spina et al. (2013) and Berrocal et al. (2013), respectively), are remarkably lower than those achieved in 2012, even taking into account the availability of training data. In any case, it seemed obvious that the topic detection problem is a complex one.

### 3.4 *Topic Ranking and Alert Detection*

Early detection of issues that may have a snowball effect is crucial for reputation management. Topics with a lot of twitter activity are more likely to have high priority. Note that experts also try to estimate how a topic will evolve in the near future. For instance, a topic may have a modest amount of tweets, but from people which are experts in the topic and have a large number of followers. A topic likely to become a trend is particularly suitable to become an alert and therefore to receive a high priority. Some of the factors that play a role in the priority assessments are:

- *Polarity*: topics with polarity (and, in particular, with negative polarity, where action is needed) usually have higher priority.
- *Centrality*: a high priority topic is very likely to have the company as the main focus of the content.
- *User's authority*: a topic promoted by an influential user (for example, in terms of the number of followers or the expertise) has better chances of receiving high priority.

Note, however, that the priority of a topic is determined by online reputation experts according to their expertise and intuitions; therefore, priority assessments will not always necessarily have a direct, predictable relationship with the factors above. This is precisely one of the issues that we wanted to investigate with this test collection.

A three-valued classification was applied to assess the priority of each entity-related topic: alert (the topic deserves immediate attention of reputation managers), mildly relevant (the topic contributes to the reputation of the entity but does not require immediate attention) and unimportant (the topic can be neglected from a reputation management perspective). Reliability represents the ratio of correct priority relationships per tweet, while Sensitivity represents the ratio of captured relationships per tweet. Results are quite similar to those achieved in the topic detection tasks, 0,27 F(R,S) for the best participant in 2012 (Martin et al. 2012) and 0,34 for the best participants in 2013 (Cossu et al. 2013).

### 3.5 *Reputational Dimension Classification*

One of the main goals when monitoring a company in Social Media is to assess the company's positioning with respect to different aspects of its activity and with respect to its peer companies. This involves a comparative analysis of the content related to that company, aiming at finding out what image the company projects in dimensions such as commercial, financial, social, labour or sectoral, and how the company's image compares to that of other companies within the same sector.

The aim of the Reputational Dimension classification in RepLab 2014 was to assign tweets to one of the seven standard reputation dimensions of the RepTrak



**Table 1** RepTrak dimensions. Definitions and examples of tweets

| Dimension             | Definition and example   |
|-----------------------|--|
| Performance           | Reflects long term business success and financial soundness of the company<br>Goldman Profit Rises but Revenue Falls: Goldman Sachs reported a second-quarter profit of \$1.05 billion, ... <a href="http://dlvr.it/bmVY4">http://dlvr.it/bmVY4</a>  |
| Products and Services | Information about the company's products and services, as well as about consumer satisfaction<br>BMW To Launch M3 and M5 In Matte Colors: Red, Blue, White but no black...   |
| Leadership            | Related to the leading position of the company<br>Goldman Sachs estimates the gross margin on ACI software to be 95% O_o   |
| Citizenship           | The company's acknowledgement of the social and environmental responsibility, including ethical aspects of business: integrity, transparency and accountability<br>Find out more about Santander Universities scholarships, grants, awards and SME Internship Programme <a href="http://bit.ly/1mM12OX">bit.ly/1mM12OX</a> |
| Governance            | Related to the relationship between the company and the public authorities<br>Judge orders Barclays to reveal names of 208 staff linked to Libor probe via @Telegraph <a href="http://soc.li/mJVPh1R">soc.li/mJVPh1R</a>   |
| Workplace             | Related to the working environment and the company's ability to attract, form and keep talented and highly qualified people<br>Goldman Sachs exec quits via open letter in The New York Times, brands bank working environment "toxic and destructive" <a href="http://ow.ly/9EaLc">ow.ly/9EaLc</a>                        |
| Innovation            | The innovativeness shown by the company, nurturing novel ideas and incorporating them into products<br>Eddy Merckx Cycles announced a partnership with Lexus to develop their ETT Hme trial bike. More info at... <a href="http://fb.me/1VAeS3zJP">http://fb.me/1VAeS3zJP</a>  |

Framework<sup>6</sup> developed by the Reputation Institute. These dimensions reflect the affective and cognitive perceptions of a company by different stakeholder groups. The task can be viewed as a complement to topic detection, as it provides a broad classification of the aspects of the company under public scrutiny. Table 1 shows the definition of each reputation dimension, supported by an example of a labelled tweet:

The system ranking for the Reputation Dimensions task was reported in terms of Accuracy. Note that tweets manually tagged as "Undefined" were excluded from the evaluation, and tweets tagged by systems as "Undefined" were considered as non-processed. The results achieved by the best team, 73% accuracy (McDonald et al. 2014), clearly outperform the proposed baseline (62% accuracy). Note that classifying every tweet in the most frequent class (majority class baseline) would

<sup>6</sup><https://www.reputationinstitute.com/about-reputation-institute/the-reprtrak-framework>.

get an accuracy of 56%. Most runs are above this threshold and provide, therefore, some useful information beyond a non-informative run.

### 3.6 *Author Classification*

The type of author may be of great interest when analysing the reputation of a company, as it may be a clear indicator of relevance. As an example, the influence of some profiles such as celebrities is of special interest for reputational experts, regardless of the domain expertise of the celebrity. The fact that the tweet author is an employee of the company, a journalist, an activist, etc., may have implications in the interpretation of the content and also in predicting its potential impact on the reputation of the entity.

The *Author Classification* task in RepLab 2014 was to classify Twitter profiles by type of author: Company (i.e., corporate accounts of the company itself), Professional (in the economic domain of the company), Celebrity, Employee, Stockholder, Investor, Journalist, Sportsman, Public Institution, and Non-Governmental Organisation (NGO). The system's output was expected to be a list of profile identifiers with the assigned categories, one per profile.

Accuracy values were computed separately for each domain (automotive, banking and miscellaneous). Average accuracy of the banking and automotive domains was used to rank systems. Interestingly, there is a high correlation between system scores in the automotive and banking domains (0,97 Pearson coefficient). The most relevant aspect of these results is that, in terms of accuracy, assigning the majority class (which is non informative) outperforms all runs (46%) except the best system (47%) (Cossu et al. 2014b). The question, then, is how much information are the systems able to produce. In order to answer this question we computed the Macro Average Accuracy (MAAC), which assigns the same (low) score to any non informative classifier. The results shows that most systems are able to improve the majority class baseline according to MAAC. This means that systems are able to abstract informative features of classes even if they make less accurate decisions than the majority class baseline.

### 3.7 *Opinion Makers Identification*

The capacity of influence of an author in the public opinion is a key element when aiming to determine the importance of topics about a company, and is the only key to fire an alert regardless of the content of the tweet. Some obvious aspects that determine the influence of an author in Twitter (from a reputation analysis perspective) are be the number of followers, number of comments on a domain or the type of author.

Using as input the same set of Twitter profiles as in the task above, systems had to find out which authors had more reputational influence (who the influencers or opinion makers are) and which profiles are less influential or have no influence at all. For a given domain (e.g., automotive or banking), systems were asked to rank profiles according to their probability of being an opinion maker in the domain, optionally including the corresponding weights. Note that, because the number of opinion makers is expected to be low, we modeled the task as a search problem (hence the system output is a ranked list) rather than as a classification problem.

The results for the Author Ranking task were ranked according to their average MAP using TREC\_EVAL software. Unfortunately, some participants returned their results in the gold standard format (binary classification as influencers or non-influencers) instead of using the prescribed ranking format. Instead of discarding those submissions, we mapped them into the official format by separating profiles marked as influencers at the top and non-influencers at the bottom of the results list, otherwise keeping the original list order.

The *followers baseline* simply ranks the authors by descending number of followers. It is clearly outperformed by most runs, indicating that additional signals provide useful information. The exception is the miscellaneous domain, where probably additional requirements over the number of followers, such as expertise in a given area, do not clearly apply. The system with the best results achieved a 0,57 MAP (McDonald et al. 2014), closely followed by Vilares et al. (2014) with a 0,56 MAP. The correlation between MAP values achieved by the systems in the automotive and banking domains seems to be low, suggesting that the performance of systems is highly biased by the domain. For future work, it is probably necessary to consider multiple domains to extract robust conclusions. On the other hand, runs from three participants exceeded 0.5 MAP, using very different approaches; Therefore, the results of the competition do not clearly point to one particular technique.

### 3.8 Full Monitoring Task

In 2013, the RepLab *full task* was a combination of all other tasks, and consisted of searching the stream of tweets for potential mentions the entity, filtering those that do refer to the entity, clustering relevant tweets by topic, and ranking topics based on their probability to be reputation alerts (i.e., issues that may have a substantial impact on the reputation of the entity, and must be handled by reputation management experts).

The use of Reliability and Sensitivity allowed us to apply the same evaluation criterion to all subtasks and therefore, to combine all of them in a single quality measure. It was possible to apply R&S directly over the full set of relationships (priority, filtering and clustering), but then the most frequent binary relationships would dominate the evaluation results (in our case, priority relationships would be predominant). Therefore, we finally decided to use a weighted harmonic mean (F

measure) of the six Reliability and Sensitivity measures corresponding to the three subtasks embedded in the full task. Due to the complex nature of this task, the results achieved by most participants were considerably low, with the best system reporting 0,19 F(R,S) (Spina et al. 2013).

This evaluation, however, is highly sensitive to the relative importance of measures in the combining function. For this reason, we also computed the Unanimous Improvement Ration (UIR) between each pair of runs. Here we considered as an unanimous improvement of system A over system B those test cases (entities) for which A improves B in all the six measures (R and S for each of the tasks). It only includes those run pairs for which UIR is bigger than 0.2. Differences in UIR turned out to be small, which indicates that the different performance of systems may not be due to intrinsic system differences, but to whether they are more optimized for reliability or sensitivity, and how this compares with the actual balance in the test data.

## 4 Post-competition Progress Using RepLab Datasets

RepLab evaluation campaigns have been, to the best of our knowledge, the most comprehensive effort to advance the understanding and automation of the online reputation management process. The availability of RepLab datasets, and the definition of the different tasks involved in the ORM process has encouraged researchers to investigate novel algorithms and methods for assisting reputational analysts in their daily work.

After the conclusion of the different RepLab editions, a good number of research teams have dealt with the problem of online ORM. Up to January 2018, RepLab overviews have received over 230 citations, and some of these citations come from studies using RepLab datasets.

In the **filtering** task, post RepLab research introduced active learning techniques to improve accuracy (Spina et al. 2015). These techniques emulate the real work of reputational analysts, interacting with the user for updating the classification model. Other recent works have employed Wikipedia to disambiguate the company's names in tweets (Qureshi et al. 2015). Others have generalized the problem of microblog filtering to consider topics of broad and dynamic nature (Magdy and Elsayed 2016).

The **reputational polarity** task has also attracted the attention of the research community after RepLab. As already mentioned, polarity for reputation strongly relies on the detection of polar facts, which is still an open problem. The most recent work known to us that has addressed the detection of polar facts in a reputational context is that of Giachanou et al. (2017), which determines the polarity of factual information by propagating the sentiment from sentiment-bearing text to factual texts that discuss the same issue. Giachanou et al. (2017) reported large improvements (over 50%) with respect to the use of sentiment analysis approaches. Previously, Peetz et al. (2016) explored the role of sender-based features (e.g., location, followers and user language), message-based features (e.g., hashtags, links

and punctuation marks) and reception-based features (e.g., sentiment strengths and scores from different lexicons). Before that, Gârbacea et al. (2014) outperformed state of the art methods using a simple supervised approach that considers three types of features: surface features (e.g., number of positive and negative words, emoticons, etc.), sentiment features (e.g., SentiWordNet scores of terms) and textual features (e.g., unigrams and bigrams). Overall, work on the RepLab dataset has clearly shown that sentiment analysis is only a starting point to deal with reputational polarity, but a lot more information is needed to provide usable results.

Post RepLab experiments in the **topic detection** task considerably improved the results of the competition. Spina et al. (2014) investigated whether it was possible to learn a generalized similarity function from the training data (to be fed in the clustering algorithm), and whether semantic signals could improve the topic detection process, with positive results in both cases. Their best system achieved a performance near inter-annotator agreement levels. They also found that the main source of disagreement was in the so-called organizational topics, while event-like topics, the ones most interesting from the point of view of reputation monitoring, were easier to handle by systems. Other approaches have employed transfer learning and LDA techniques by contextualizing a target collection of tweets with a large set of unlabeled “background” tweets (Martín-Wanton et al. 2013). In Panem et al. (2014), two unsupervised approaches are presented, the first based on keyword extraction and keyphrase identification, and the second based on a conceptual representation using Wikipedia.

The **priority** task has only attracted limited attention from the research community after RepLab. Cossu et al. (2014a) presented the only work that, to the best of our knowledge, has addressed the problem after the RepLab campaigns. They combine different clustering for topic detection with different priority classification methods, and conclude that actual methods are not yet mature enough to reach better performances than any priority assignment system taken alone.

With respect to **author profiling**, post-RepLab research has focused on the study of Twitter features that are relevant to characterize influential profiles (Cossu et al. 2014b), including features related to the user activity, the network topology, stylistic aspects, tweets characteristics, and profile fields. Mabrouk et al. (2018) proposed a simple model based on tf\*idf and feature vector reduction. Mahalakshmi et al. (2017) propose to find the influential users in a community using a combination of the user position in networks that emerge from Twitter relations, and the textual quality of her tweets. Nebot et al. (2018) experimented with deep neural networks and word embeddings obtaining competitive results; and recently, Rodriguez et al. (2019b) investigated the different roles of authority signals (those that point out that the user is an influencer) and domain signals (those that indicate that the user is associated with the economic domain of interest) in detecting domain-specific opinion makers, and found out that both can be handled effectively with language models of influencers in the domain. Both in Nebot et al. (2018) and Rodriguez et al. (2019b), one of the salient conclusions is that text contains enough information to address the task, and additional non-textual signals, which in principle seem very

relevant for the problem, such as the number of followers, do not improve the use of textual information.

As for the task of **classification into reputational dimensions**, Qureshi et al. (2017) obtain Wikipedia dominant categories to generate “associativeness” with respect to the various reputation dimensions, and then are used in a random forest classifier, showing significant improvement over the baseline accuracy. McDonald et al. (2015) present a tweet enrichment approach that expands tweets with additional discriminative terms from a contemporary Web corpus, and that outperforms effective baselines including the top performing submission to RepLab 2014.

Work on RepLab data goes beyond the tasks defined in the evaluation exercise. A new and strongly related task has emerged post-RepLab: the automatic generation of reputational reports using the output of the tasks investigated in RepLab. Carrillo-de Albornoz et al. (2016) investigated the problem with two goals: determining if it is substantially different from a standard summarization task, and finding out appropriate evaluation metrics. Their experiments showed that producing reputation reports differs from standard summarization in the key role played by the reputational priority of information nuggets, which must be handled by systems together with centrality (the standard signal in summarization). In Rodriguez et al. (2019a), a test collection for the task of producing reputation reports is created, with extractive and abstractive summaries manually created for each of the alerts and important topics identified in each of the RepLab 2013 entities.

Finally, it is worth mentioning that the websites of the competitions have been accessed over 8000 times by more than 5000 different users, and that the datasets and results of the different systems are available for the research community in the EvALL (Amigó et al. 2017) framework (<http://evall.uned.es/>).

## 5 Discussion

Over a period of 3 years, RepLab was a CLEF Lab where computer scientists and online reputation experts worked together to identify and formalize the Natural Language Processing challenges in the area of online reputation monitoring. Two main results emerged from RepLab: a community of researchers engaged in the problem, and an extensive Twitter test collection comprising more than half a million expert annotations covering many relevant tasks in the field of online reputation: named entity resolution, topic detection and tracking, reputational alerts identification, reputational polarity, author profiling, opinion makers identification and reputational dimension classification. It has probably been the CLEF lab with the largest set of expert annotations provided to participants in a single year, and one of the labs where a user community has been more actively engaged in an evaluation initiative. Four years after completion of the lab, RepLab data is still being used by the research community.

A characteristic of the problems studied in RepLab is the size of the data to be handled per client: in a typical case, online information about a company is too much

to be processed manually, but too little to apply the simple statistics that are perfectly fit for massive trending topics. Companies are, so to speak, in the “long tail” of social media information, except for a handful of prominent multinational corporations such as Coca Cola, Apple, etc. Another key feature of dealing with reputation monitoring is that straightforward Machine Learning is usually not enough; the focus of reputation monitoring is on early discovery of the unexpected (an issue that arises about the entity that was not foreseen). And, from that point of view, a Machine Learning algorithm has to be able to generalize in a very clever way to distinguish a new reputational issue based on what has been seen and tagged before. Often, Machine Learning methods extract statistics from data that do not generalize well on new material; for instance, they can learn that “ecologist” is a term that usually correlates with something bad for the reputation of oil companies; if the unseen data unexpectedly contains some positive actions of oil companies in the environment, the algorithm will fail to analyze that content properly.

The close work with reputation experts did not stop at RepLab; in the framework of the Limosine project which funded the evaluation campaigns, the consortium built and tested annotation assistants with the help of the experts. There, we discovered that the main scientific findings in RepLab did not necessarily correlate to the techniques needed to optimize the work of the experts. For instance, in Spina et al. (2014) we discovered that semantic signals (such as entity linking of tweet terms with Wikipedia entries) could improve topic detection in a statistically significant way. In practice, however, it was preferable to deploy a system able to re-train very fast when the experts corrected an automatic detection; fast adaptive learning was far more important than the level of sophistication of the signals used for the initial automatic annotation.

**Acknowledgement** This research was partially supported by the Spanish Ministry of Science and Innovation (Vemodalen Project, TIN2015-71785-R).

## References

- Alsop RJ (2006) The 18 immutable laws of corporate reputation: creating, protecting and repairing your most valuable asset. Kogan Page, London
- Amigó E, Gonzalo J, Artiles J, Verdejo F (2009) A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf Retr* 12(4):461–486
- Amigó E, Gonzalo J, Artiles J, Verdejo F (2011) Combining evaluation metrics via the unanimous improvement ratio and its application to clustering tasks. *J Artif Intell Res* 42(1):689–718
- Amigó E, Gonzalo J, Verdejo F (2013) A general evaluation measure for document organization tasks. In: Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, SIGIR '13, pp 643–652. <https://doi.org/10.1145/2484028.2484081>
- Amigó E, Carrillo-de Albornoz J, Almagro-Cádiz M, Gonzalo J, Rodríguez-Vidal J, Verdejo F (2017) Evall: open access evaluation for information access systems. In: Proceedings of the 40th international ACM SIGIR conference on research and development in information

- retrieval. ACM, New York, NY, SIGIR '17, pp 1301–1304. <https://doi.org/10.1145/3077136.3084145>
- Balahur A, Tanev H (2012) Detecting entity-related events and sentiments from tweets using multilingual resources. In: CLEF (Online Working Notes/Labs/Workshop)
- Berrocal A, Luis J, Figuerola CG, Zazo Rodríguez ÁF (2013) Reina at replab2013 topic detection task: community detection. In: CLEF (Working Notes)
- Carrillo-de-Albornoz J, Chugur I, Amigó E (2012) Using an emotion-based model and sentiment analysis techniques to classify polarity for reputation. In: CLEF (Online Working Notes/Labs/Workshop)
- Carrillo-de Albornoz J, Amigó E, Plaza L, Gonzalo J (2016) Tweet stream summarization for online reputation management. In: Ferro N, Crestani F, Moens MF, Mothe J, Silvestri F, Di Nunzio GM, Hauff C, Silvello G (eds) *Advances in information retrieval*. Springer International Publishing, Cham, pp 378–389
- Chong WN, Tan G (2010) Obtaining intangible and tangible benefits from corporate social responsibility. *Int Rev Bus Res Pap* 6(4):360
- Cossu JV, Bigot B, Bonnefoy L, Morchid M, Bost X, Senay G, Dufour R, Bouvier V, Torres-Moreno JM, El-Bèze M (2013) Lia at replab 2013. In: CLEF (Working Notes)
- Cossu JV, Bigot B, Bonnefoy L, Senay G (2014a) Towards the improvement of topic priority assignment using various topic detection methods for e-reputation monitoring on twitter. In: Métais E, Roche M, Teisseire M (eds) *Natural language processing and information systems*. Springer International Publishing, Cham, pp 154–159
- Cossu JV, Janod K, Ferreira E, Gaillard J, El-Bèze M (2014b) Lia at replab 2014: 10 methods for 3 tasks. In: 4th international conference of the CLEF initiative
- Doorley J, Garcia HF (2011) *Reputation management: the key to successful public relations and corporate communication*. Routledge, New York
- Filgueiras J, Amir S (2013) Popstar at replab 2013: polarity for reputation classification. In: CLEF (Online Working Notes/Labs/Workshop), vol 60
- Gârbacea C, Tsagkias M, de Rijke M (2014) Detecting the reputation polarity of microblog posts. In: *Proceedings of the twenty-first european conference on artificial intelligence*. IOS Press, Amsterdam, ECAI'14, pp 339–344. <https://doi.org/10.3233/978-1-61499-419-0-339>
- Giachanou A, Gonzalo J, Mele I, Crestani F (2017) Sentiment propagation for predicting reputation polarity. In: Jose JM, Hauff C, Altingovde IS, Song D, Albakour D, Watt S, Tait J (eds) *Advances in information retrieval*. Springer International Publishing, Cham, pp 226–238
- Glance N, Hurst M, Nigam K, Siegler M, Stockton R, Tomokiyo T (2005) Deriving marketing intelligence from online discussion. In: *Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining*. ACM, New York, NY, KDD '05, pp 419–428. <https://doi.org/10.1145/1081870.1081919>
- Hangya V, Farkas R (2013) Filtering and polarity detection for reputation management on tweets. In: CLEF (Online Working Notes/Labs/Workshop), vol 60
- Hoffman T (2008) Online reputation management is hot – but is it ethical. *Computerworld* (44). <https://www.computerworld.com/article/2537007/networking/online-reputation-management-is-hot---but-is-it-ethical-.html>
- Jansen B, Zhang M, Sobel K, Chowdury A (2009a) Twitter power: tweets as electronic word of mouth. *J Am Soc Inf Sci Technol* 60(11):2169–2188
- Jansen BJ, Zhang M, Sobel K, Chowdury A (2009b) Twitter power: tweets as electronic word of mouth. *J Assoc Inf Sci Technol* 60(11):2169–2188
- Kaptein R (2012) Learning to analyze relevancy and polarity of tweets. In: CLEF (Online Working Notes/Labs/Workshop), vol 60
- Kreps DM, Wilson R (1982) Reputation and imperfect information. *J Econ Theory* 27(2):253–279
- Krishnamurthy B, Gill P, Arlitt M (2008) A few chirps about Twitter. In: *Proceedings of the first workshop on online social networks (WOSP'08)*, pp 19–24
- Li YM, Li TY (2013) Deriving market intelligence from microblogs. *Decis Support Syst* 55(1):206–217. <https://doi.org/10.1016/j.dss.2013.01.023>. <http://www.sciencedirect.com/science/article/pii/S0167923613000511>



- Mabrouk O, Hlaoua L, Nazih Omri M (2018) Profile categorization system based on features reduction. In: International symposium on artificial intelligence and mathematics, ISAIM 2018, Fort Lauderdale, Florida. [http://isaim2018.cs.virginia.edu/papers/ISAIM2018\\_Mabrouk\\_et\\_al.pdf](http://isaim2018.cs.virginia.edu/papers/ISAIM2018_Mabrouk_et_al.pdf)
- Magdy W, Elsayed T (2016) Unsupervised adaptive microblog filtering for broad dynamic topics. *Inf Process Manag* 52(4):513–528
- Mahalakshmi GS, Koquilamballe K, Sendhilkumar S (2017) Influential detection in twitter using tweet quality analysis. In: 2017 second international conference on recent trends and challenges in computational models (ICRTCCM), pp 315–319. <https://doi.org/10.1109/ICRTCCM.2017.62>
- Martin T, Spina D, Amigó E, Gonzalo J (2012) Uned at replab 2012: monitoring task. In: CLEF 2012 Working Notes, CLEF
- Martin-Wanton T, Gonzalo J, Amigó E (2013) An unsupervised transfer learning approach to discover topics for online reputation management. In: Proceedings of the 22nd ACM international conference on conference on information & knowledge management, ACM, New York, NY, CIKM '13, pp 1565–1568. <https://doi.org/10.1145/2505515.2507845>
- McDonald G, Deveaud R, McCreddie R, Gollins T, Macdonald C, Ounis I (2014) University of glasgow terrier team/project abacá at replab 2014: reputation dimensions task. In: CLEF (Working Notes), pp 1500–1504
- McDonald G, Deveaud R, McCreddie R, Macdonald C, Ounis I (2015) Tweet enrichment for effective dimensions classification in online reputation management. In: 9th international AAAI conference on web and social media, pp 654–657
- Nebot V, Rangel F, Berlanga R, Rosso P (2018) Identifying and classifying influencers in twitter only with textual information. In: Proceedings of the NLDB 2018
- Panem S, Bansal R, Gupta M, Varma V (2014) Entity tracking in real-time using sub-topic detection on twitter. In: de Rijke M, Kenter T, de Vries AP, Zhai C, de Jong F, Radinsky K, Hofmann K (eds) *Advances in information retrieval*. Springer International Publishing, Cham, pp 528–533
- Peetz MH, de Rijke M, Kaptein R (2016) Estimating reputation polarity on microblog posts. *Inf Process Manag* 52(2):193–216. <https://doi.org/10.1016/j.ipm.2015.07.003>. <http://www.sciencedirect.com/science/article/pii/S0306457315000874>
- Qureshi M, Younus A, O’Riordan C, Pasi G (2015) Company name disambiguation in tweets: a two-step filtering approach. In: *Information retrieval technology*, vol 9460
- Qureshi MA, O’Riordan C, Pasi G (2012) Concept term expansion approach for monitoring reputation of companies on twitter. In: CLEF (Online Working Notes/Labs/Workshop)
- Qureshi MA, Younus A, O’Riordan C, Pasi G (2017) A wikipedia-based semantic relatedness framework for effective dimensions classification in online reputation management. *J Ambient Intell Humaniz Comput* 9:1403
- Rodriguez J, Carrillo-de Albornoz J, Plaza L, Amigó E, Gonzalo J (2019a) Automatic generation of entity-oriented summaries for reputation management. *J Ambient Intell Humaniz Comput*:1–15. <https://link.springer.com/article/10.1007/s12652-019-01255-9>
- Rodriguez J, Gonzalo J, Plaza L, Anaya H (2019b) Automatic detection of influencers in social networks: authority versus domain signals. *J Assoc Inf Sci Technol* 70:7
- Spina D, Carrillo-de-Albornoz J, Martín-Wanton T, Amigó E, Gonzalo J, Giner F (2013) Uned online reputation monitoring team at replab 2013. In: CLEF (Working Notes)
- Spina D, Gonzalo J, Amigó E (2014) Learning similarity functions for topic detection in online reputation monitoring. In: Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval. ACM, New York, NY, SIGIR '14, pp 527–536. <https://doi.org/10.1145/2600428.2609621>
- Spina D, Peetz MH, de Rijke M (2015) Active learning for entity filtering in microblog streams. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, SIGIR '15, pp 975–978. <https://doi.org/10.1145/2766462.2767839>

- Vilares D, Hermo M, Alonso MA, Gómez-Rodríguez C, Vilares J (2014) Lys at clef replab 2014: creating the state of the art in author influence ranking and reputation classification on twitter. In: CLEF (Working Notes), pp 1468–1478
- Villena-Román J, Lana-Serrano S, Moreno C, García-Morera J, Cristóbal JCG (2012) Daedalus at replab 2012: polarity classification and filtering on twitter data. In: CLEF (Online Working Notes/Labs/Workshop), vol 60

# Continuous Evaluation of Large-Scale Information Access Systems: A Case for Living Labs



Frank Hopfgartner, Krisztian Balog, Andreas Lommatzsch, Liadh Kelly, Benjamin Kille, Anne Schuth, and Martha Larson

**Abstract** A/B testing is currently being increasingly adopted for the evaluation of commercial information access systems with a large user base since it provides the advantage of observing the efficiency and effectiveness of information access systems under real conditions. Unfortunately, unless university-based researchers closely collaborate with industry or develop their own infrastructure or user base, they cannot validate their ideas in live settings with real users. Without online testing opportunities open to the research communities, academic researchers are unable to employ online evaluation on a larger scale. This means that they do not get feedback for their ideas and cannot advance their research further. Businesses, on the other hand, miss the opportunity to have higher customer satisfaction due to improved systems. In addition, users miss the chance to benefit from an improved information access system. In this chapter, we introduce two evaluation initiatives at CLEF,

---

F. Hopfgartner (✉)  
University of Sheffield, Sheffield, UK  
e-mail: [f.hopfgartner@sheffield.ac.uk](mailto:f.hopfgartner@sheffield.ac.uk)

K. Balog  
University of Stavanger, Stavanger, Norway  
e-mail: [krisztian.balog@uis.no](mailto:krisztian.balog@uis.no)

A. Lommatzsch · B. Kille  
Technische Universität Berlin, Berlin, Germany  
e-mail: [andreas.lommatzsch@dai-labor.de](mailto:andreas.lommatzsch@dai-labor.de); [benjamin.kille@dai-labor.de](mailto:benjamin.kille@dai-labor.de)

L. Kelly  
Maynooth University, Maynooth, Ireland  
e-mail: [liadh.kelly@mu.ie](mailto:liadh.kelly@mu.ie)

A. Schuth  
De Persgroep, Amsterdam, The Netherlands

M. Larson  
Radboud University, Nijmegen, The Netherlands  
e-mail: [m.larson@cs.ru.nl](mailto:m.larson@cs.ru.nl)

NewsREEL and Living Labs for IR (LL4IR), that aim to address this growing “evaluation gap” between academia and industry. We explain the challenges and discuss the experiences organizing these living labs.

## 1 Introduction

As evident from the other chapters of this book, significant efforts have been invested in establishing metrics, frameworks, and datasets to guarantee a thorough and transparent evaluation of novel approaches to retrieve or recommend documents and items. For many years, campaigns such as CLEF, TREC, NTCIR, and FIRE have played leading roles in promoting research in the field of information retrieval. The release of datasets, standardized evaluation metrics and evaluation procedures, following the established Cranfield evaluation paradigm, has contributed to innovative retrieval approach development in domains such as newswire articles, blogs, microblogs, and biomedical documents to name but a few. In the field of recommender systems research, a similar coordinated evaluation procedure with standardized datasets and evaluation criteria has been established thanks to the release of the Netflix dataset and the associated challenge, as well as the release of the MovieLens datasets. In both cases, it is safe to claim that the release of test collections was of great benefit for the research community since it spared researchers not only from the tedious task of creating their own datasets, but also allowed them to easily compare their results with state-of-the-art algorithms. However, as Voorhees and Harman (2005) point out, the use of standardized datasets also comes with certain drawbacks. In many research papers, datasets are used to fine-tune computational models or algorithms, resulting in improved performance, e.g., measured based on precision, recall, or using other popular metrics. This is a direct consequence of the ability to compare performance against state-of-the-art approaches and the desire to beat those baselines.

This limitation is well understood by commercial providers of information access systems who rely increasingly on user-centric evaluation of their systems to achieve optimal performance (Kohavi 2015). The large number of users of their systems implicitly allows for evaluation of the efficiency and effectiveness of algorithms under real conditions as they engage with the systems. This has resulted in this user-centric evaluation paradigm evolving into the de-facto evaluation standard employed in commercial settings. Evaluation of this nature is referred to as *online evaluation* since it is employed using instances of online information access systems, or as *A/B testing* since it allows for the comparison of different variants of the system. Unfortunately, non-commercial, especially university-based, researchers are now struggling to evaluate their own approaches using this resource-demanding evaluation standard. This was also pointed out by Hawking (2015) who compared the affiliation of authors’ of research papers presented at SIGIR’98 and SIGIR’15, respectively. He argued that the observed increase from 15% of industrial research papers published in 1998 compared to 41% published in 2015 is a direct consequence of the increased need to evaluate research methods using large-scale datasets or user studies.

Addressing the lack of access to data, Hanbury et al. (2015) argue for the implementation of evaluation services that store data on a central server and allow researchers access to both data and information technology infrastructure. They refer to this method as Evaluation-as-a-Service (EaaS). While this approach has the potential to alleviate the growing evaluation gap to some extent, it does not address the issue of having limited access to real users who can be test subjects for researchers' algorithms and ideas. To address this, the application of a living lab that grants researchers access to real users who follow their own information seeking tasks in a natural and thus realistic contextual setting has been proposed (Kamps et al. 2009; Kelly et al. 2009). For user-centric research on information access systems, realistic context is essential since it is a requirement for a fair and unbiased evaluation. In this chapter, we present the two living labs initiatives that have been introduced within the domains of recommender systems and information retrieval (IR).

The CLEF NewsREEL challenge is a campaign-style evaluation lab allowing participants to evaluate and optimize news recommender algorithms. The goal is to create an algorithm that is able to generate news items that users would click on, respecting a strict time constraint. The lab challenged participants to compete in either a living lab or perform an evaluation that replays recorded streams. By participating in this living lab, participants are given the opportunity to develop news recommendation algorithms and have them tested by potentially millions of users of a live system over a longer period of time.

The Living Labs for Information Retrieval (LL4IR) CLEF lab is a benchmarking platform for researchers to evaluate their retrieval systems in a live setting. The lab acts as a proxy between commercial organizations (live environments) and lab participants (experimental systems), facilitates data exchange, and makes comparison between the participating systems possible. The LL4IR lab focused on two use cases: product search (on an e-commerce site) and web search (through a commercial web search engine).

After surveying state-of-the-art in the area of online evaluation in Sect. 2, we present the NewsREEL (Sect. 3) and LL4IR (Sect. 4) use cases as leading examples of living labs evaluation. In Sect. 5 we highlight similarities and differences between the two approaches, and conclude with a discussion on the opportunities and challenges that such online evaluation campaigns offer.

## 2 Related Work

Information access systems have been evaluated in four major ways (Gunawardana and Shani 2009): offline with static test collections, with small-scale user studies or user simulations, and in online evaluation environments. Tradition has favored offline evaluation to ensure reproducibility. At the same time, such evaluation may not accurately reflect user satisfaction (Teevan et al. 2007; Turpin and Scholar 2006; Wilkins et al. 2008). Moreover, it leaves one of the most important factors of any

information retrieval or recommender system out of the loop: the user. It is the user's information need that needs to be satisfied and it is the user's personal interests that need to be considered when providing personalized access to information. This is one of the major reasons for performing *online* evaluation: evaluation with users in the loop.

The need for more realistic evaluation, involving real users, has been reiterated at several workshops (Kamps et al. 2009; Allan et al. 2012; Balog et al. 2014a). To address this, living labs have emerged as a way for researchers to be able to perform *in situ* evaluation. The main idea behind living labs is that an existing information access service serves as the experimentation platform. By replacing components of this information access platform, researchers have the opportunity to perform evaluation using interactions with real, unsuspecting users of this information access system. Major information access online evaluations and A/B testing are instances of living labs. However, this type of evaluation has only recently become available to the broader research community.

## 2.1 *Living Labs Shared Challenges*

The notion of using living labs for shared challenges in the information access space has been proposed in recent years (Azzopardi and Balog 2011; Kelly et al. 2012). In particular, Azzopardi and Balog (2011) present details on an approach to move from a traditional IR evaluation setting to a living labs setting. The first implementation of a living lab was the NewsREEL challenge that was first organized as part of a workshop co-located with ACM RecSys (Tavakolifard et al. 2013). Later, it was operated as part of CLEF. NewsREEL allowed participants to evaluate and optimize news recommendation algorithms. The goal was to create an algorithm for news recommendation that is able to generate news items that users would click on, respecting a strict time constraint for generating and serving those recommendations. By participating in NewsREEL, researchers who develop stream-based recommendation algorithms could have these benchmarked by actual users of a live system over a longer period of time (Hopfgartner et al. 2015a). In the context of information retrieval, Balog et al. (2014b) proposed a practical way of operationalizing the living lab idea by limiting evaluation to head queries, a setup that was subsequently adopted by the CLEF LL4IR lab (cf. Sect. 4.1). The same idea was also employed at the TREC 2016 and 2017 OpenSearch track, where the use case is scientific literature search (Jagerman et al. 2018). Kelly et al. (2012) presented an alternative living labs setting as a solution to the evaluation of personal search.

## 2.2 Online Testing

*A/B Testing* compares two systems by showing system A to one group of users and system B to a disjoint group (Kohavi 2015). The difference between the systems is inferred from observed user behavior. This includes, among other things, click-through rate (CTR) (Joachims et al. 2007), dwell time (Yilmaz et al. 2014), satisfied clicks (Kim et al. 2014), abandonment (Li et al. 2009), query reformulation (Hassan et al. 2013), and mouse movement (Wang et al. 2010; Diaz et al. 2013). NewsREEL, for example, employed the click-through rate as its primary evaluation criterion.

An alternative to A/B testing is to perform interleaved comparisons, which are shown to be more sensitive (Schuth et al. 2015c; Chapelle et al. 2012). This means that far fewer query impressions are required to make informed decisions on which ranker is better. Many interleaving approaches have been proposed over the past few years, see, e.g., Joachims (2003); Radlinski et al. (2008); Hofmann et al. (2011); Radlinski and Craswell (2013); Schuth et al. (2014, 2015b). By far the most frequently used interleaving algorithm to date is Team Draft Interleaving (TDI) (Radlinski et al. 2008) which is also what is used in the CLEF LL4IR lab. Given a user query  $q$ , TDI produces an interleaved result list as follows. The algorithm takes as input two rankings. One ranking from the participant  $r' = (a_1, a_2, \dots)$  and one from the production system  $r = (b_1, b_2, \dots)$ . The goal is to produce a combined, interleaved ranking  $L = (a_1, b_2, \dots)$ . This is done similarly to how sports teams may be constructed in a friendly sports match. The two team captains take turns picking players. They can pick available documents (players) from the top of the rankings  $r'$  and  $r$ , these top ranked documents are deemed to be the best documents. Documents can only be picked once (even if they are listed in both  $r$  and  $r'$ ). And the order in which the documents are picked determines ranking  $L$ . In each round, the team captains flip a coin to determine who goes first. The algorithm remembers which team each document belongs to. If a document receives a click from a user, credit is assigned to the team the document belongs to. The team (participant or production system) with most credit wins the interleaved comparison. This process is repeated for each query. For more details see the original paper describing TDI by Radlinski et al. (2008) and a large-scale comparison of interleaving methods by Chapelle et al. (2012).

## 3 News Recommendation Evaluation Lab (NewsREEL)

The first information access living lab that is introduced in this chapter focuses on the domain of news recommendation. Recommender systems pro-actively suggest information to users based on their preferences. The first recommender systems entered the realm of online content distribution in the 2000s. Unfortunately though, after a decade of research, a gap emerged between academia and

industry. Academia focused on experimenting with fixed datasets often neglecting practical aspects of recommender systems. Industry, on the other hand, implemented A/B testing procedures. As discussed in Sect. 2.2, this procedure partitions users into groups, exposes them to variations of the system, and monitors differences in performance. While academia achieved repeatability of experiments, industry observes the actual reactions of users. NewsREEL, short for News Recommendation Evaluation Lab, was a campaign style evaluation task designed to bridge this gap.<sup>1</sup> It was first organized in conjunction with the ACM RecSys 2013 Workshop on News Recommender Systems (Tavakolifard et al. 2013) and then joined CLEF as campaign-style lab between 2014 and 2017. The four CLEF editions observed a total of 230 registrations. NewsREEL afforded participants the opportunity to engage in both offline and online evaluations. On the one hand, participants had access to a large-scale stream of recorded events, which could be used for offline comparison of different algorithms. On the other hand, participants gained access to a commercial news recommender system which delivered suggestions for a set of publishers in real-time. This provided participants with access to authentic live recommender system conditions. Developing recommender services in this environment represents a challenging task. Challenges included overcoming issues of availability, responsiveness, and scalability beside algorithmic design and optimization. In particular, the environment is subject to change. Publishers push new articles as events happen. Readers' interests shift over time. Hence, models have to be updated.

In the remainder of this section, we first describe the news recommendation problem addressed by NewsREEL and introduce the online and offline tasks, NewsREEL Live and NewsREEL Replay (Sect. 3.1). While the online task requires participants to provide recommendations to real users in real-time, the offline task can be run on standalone hardware without online access and the necessity to fulfill specific time constraints. In addition, the offline task simplifies the debugging and the simulation of streams. Algorithms shown to be working offline can then be evaluated in the NewsREEL Live task without any changes. Section 3.2 describes the NewsREEL evaluation architecture. We discuss participation in the online challenge in Sect. 3.3 and the offline challenge in Sect. 3.4. Section 3.5 provides a discussion on NewsREEL.

### 3.1 *NewsREEL Use Case*

As previously mentioned, CLEF NewsREEL implemented a shared challenge in the news recommendation space. It consisted of two tasks that were based on the use case of providing a list of news articles relevant to a given new article that a reader might be interested in. As depicted in Fig. 1, these news article recommendations are

---

<sup>1</sup>See <http://newsreelchallenge.org/> for details.



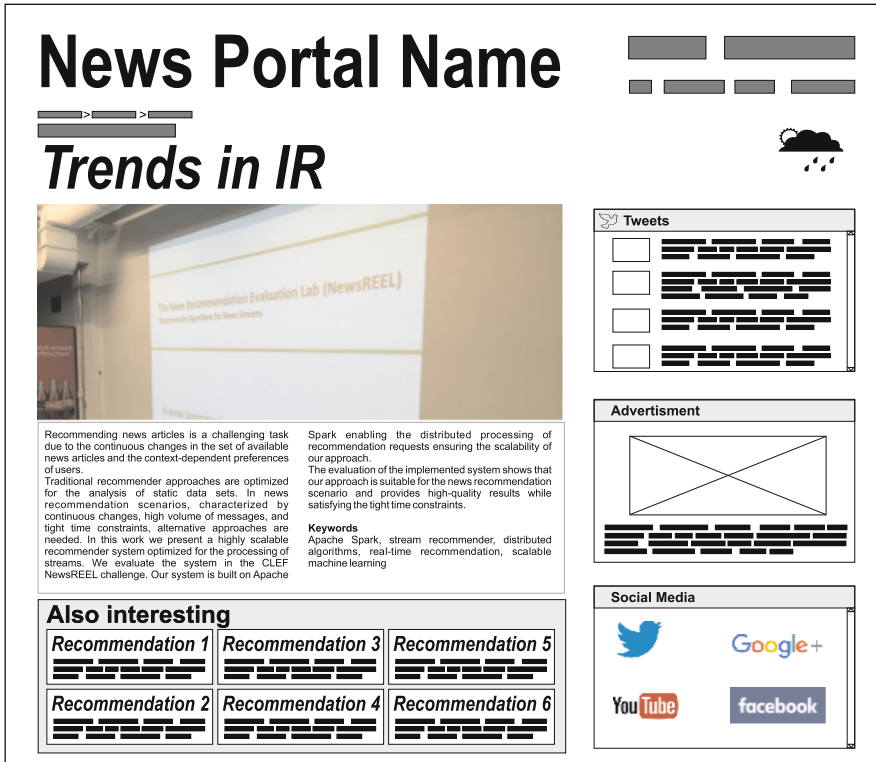


Fig. 1 Exemplary illustration of the way news recommendations are displayed to readers in the NewsREEL scenario

often displayed at the bottom or the side of the article. Determining what articles to suggest to readers is challenging from a technical point of view. First of all, recommendations have to be displayed to readers in real-time. Moreover, publishers have relatively limited information about readers and their interests. Supply and demand of information are continuously subject to change. Besides, publishers constantly add new articles and readers may lose interest in events or move on to different topics. News recommender systems have to adapt to these dynamics. The two tasks are outlined in detail in the remainder of this section. For a more detailed description of the NewsREEL use case, we refer the reader to Hopfgartner et al. (2015a).

### 3.1.1 Online Evaluation of News Recommendation Algorithms

The first NewsREEL task implemented a living lab style shared challenge. This living labs evaluation challenge is described in detail in Hopfgartner et al. (2014). Researchers gained access to resources of the online information service provider

plista<sup>2</sup> such that they could conduct A/B testing for a selection of recommendation techniques. Plista offers recommendation services and targeted advertisements for online publishers. As users request articles from publishers' web portals, plista provides a list of additional suggested articles. Plista forwards a random subset of these request to NewsREEL's participants via the Open Recommendation Platform (ORP) (see Brodt and Hopfgartner 2014). In addition, participants received information about the overall activity on the publishers' platform in the form of reads, clicks on suggestions, as well as new or updated articles. Participants needed to respond to requests within 100 ms.

### 3.1.2 Offline Evaluation of News Recommendation Algorithms

The second task addressed the academic perspective of focusing on reproducibility of results. Tools to replay the event stream allowed participants to compare algorithms and parameter configurations in identical conditions. In addition, participants could determine time and space complexity of their algorithms. Kille et al. (2015) describe the offline task in greater detail.

We have released multiple large datasets comprising interactions between users and articles on various publishers sites. The datasets' characteristics are described in detail in Kille et al. (2013). The news portals publish mostly German articles. Consequently 80% of readers reside in the German-speaking area of Central Europe (Germany, Austria, and Switzerland). Figure 2 illustrates the geographical spread of user activity. Moreover, we have released a toolkit called idomaar (Scriminaci et al. 2016) that allowed participants to "replay" the dataset.

## 3.2 NewsREEL Architecture

NewsREEL has been designed with reusability in mind. Both tasks assessed the quality of recommendation strategies for news. In the online living labs task implicit feedback was received from users of the live publishers sites. The offline task estimated relative quality on a recorded stream of event messages. The tasks shared a common interface for recommendation algorithms. Thus, participants could deploy their algorithms in both tasks without additional costs. In the online task, the ORP handled communication and monitoring of feedback. In the offline task, a replaying service took the recorded streams as input, issued requests to the algorithms being evaluated, and kept track of the results. Figure 3 depicts the NewsREEL architecture. In both settings, requests emerged, were forwarded to a recommender, suggestions were delivered, and their performance was assessed. In the offline task, the contest server delivered a summary of the response times. This lets participants judge whether the algorithm is suited for online deployment. In the

---

<sup>2</sup><http://plista.com/>.

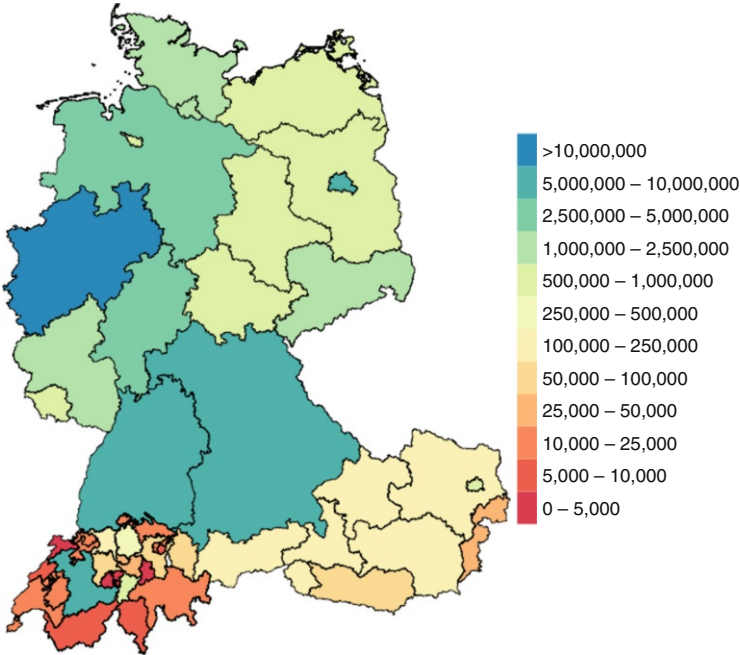


Fig. 2 Areas in Germany, Austria, and Switzerland from where requests for articles were triggered. The scale indicates the number of requests during 1 month

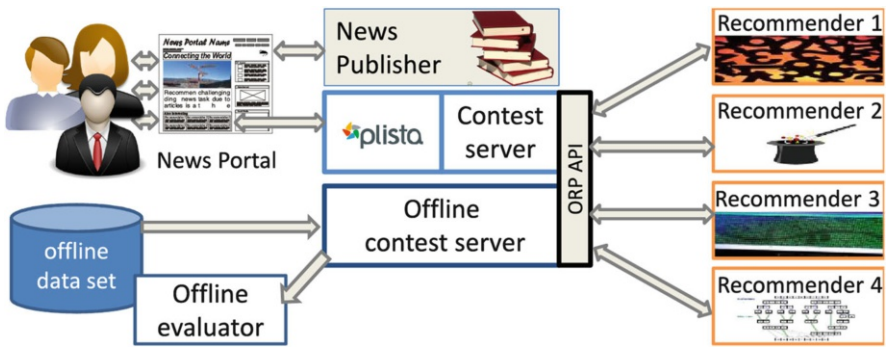


Fig. 3 The NewsREEL challenge architecture

online task, ORP ignored recommendations arriving outside the defined response time limit. Thus, the more algorithms exceeded this threshold, the more the click-through rate decreased. In both settings, communication was based on HTTP. Data are exchanged in JSON format. Interfacing with publishers and providing large-scale data collections, NewsREEL represented a unique opportunity for academic researchers to experience a setting close to the industrial reality.

### 3.3 NewsREEL Recommender Algorithms

In the NewsREEL challenge, participants evaluated a wide spectrum of recommender approaches. In this section, we briefly summarize trialled methods and discuss their relation to the living labs environment. A more detailed overview of the strengths and limitations of these methods is currently under preparation.

#### 3.3.1 The Algorithms Evaluation in the Online Task

*Big Data Frameworks* Rapidly changing user preferences and strict requirements with respect to scalability and response time represented a major challenge for NewsREEL's participants. Several authors used big data frameworks to fulfill these requirements. Verbitskiy et al. (2015) developed a most-popular recommender using the AKKA framework benefiting from concurrent message passing. They registered a high click-through rate while simultaneously ensuring fast responses.

Ciobanu and Lommatzsch (2016) developed a stream-based news recommender using APACHE FLINK. They performed well even though the systems suffered from breaking streams in the long-term evaluation.

Several authors (Lommatzsch et al. 2016; Domann et al. 2016; Beck et al. 2017) have used APACHE SPARK and APACHE MAHOUT. The combinations facilitate periodically building new micro-batches to update the models. All these approaches outperformed the baseline while ensuring high scalability.

*Graph- and Rule-Based Approaches* Bons et al. (2017) developed a graph-based recommender algorithm. The graph consisted of nodes representing the items and directed edges describing the frequency and sequence in which the two connected news items were read. Recommendation requests were answered by computing the strongest item sequence containing the itemID given in the recommendation request. The graph was managed in a Neo4j graph database. Recommendations were computed based on a database query. If the itemID in the recommendation request did not exist in the graph or the node was not yet connected with the graph, the most recently created news items were returned. The evaluation of the strategy showed that the implemented graph-based recommender reached a high click-through rate in the Living Labs scenario. The implementation worked efficiently, ensuring that the time-constraints with respect to response time were reliably fulfilled.

Golian and Kuchar (2017) analyzed click patterns in time series from NewsREEL 2016. They showed that a limited set of news items attract a majority of clicks, and that they continue to dominate for longer times than expected. They conducted a series of experiments in the context of online news recommender system evaluation. The authors report that content-based methods achieve considerably lesser click-through rates than popularity-based methods.

Ludmann (2017) focused on managing streams. His system relied on *Odysseus*, a data stream management system. He defined a set of queries which took parts of the data stream and determined the most popular articles. The selection entailed

the length of the data stream segment as essential parameter. They presented observations on NewsREEL Live with a variety of parameter configurations. Results suggest that considering successful recommendations improves the click-through rates.

*Recommender Ensembles* The continuous changes in the data stream motivated several participants to implement an ensemble recommender. Beck et al. (2017) used an ensemble of a user-based collaborative (CF) and a most popular (“unpersonalized”) recommender. The CF-based recommender provided personalized recommendation for users with session-profiles. The most popular recommender provided recommendations for new users (overcoming the cold-start problem). More complex ensembles combining different content-based and CF-based recommender algorithms are presented in Lommatzsch and Albayrak (2015). The developed system estimated the performance of the different recommender algorithms in different contexts (defined by on time and type of recommendation requests). The system learned which algorithms performed best for each context—new requests were delegated to the most promising algorithm. The ensemble approach outperformed all teams using only a single algorithm.

Gebremeskel and de Vries (2015) explored the utility of geographic information. They hypothesized that visitors have special interest in news stories about their local community. They implemented a recommender which leveraged geographic data when matching visitors and news articles.

Corsini and Larson (2016) discussed how images affect users’ response to recommendations. They argued that selecting promising images increases the likelihood of clicks. They introduced an image processing pipeline. The pipeline detects faces and image salience. A binary classifier subsequently decided whether an image is interesting or not. The authors evaluated the approach offline and online. They report improvements in the offline case. Further work is necessary to achieve reliable online evaluation results.

Liang et al. (2017) discussed how contextual bandits can be used to compute recommendations. The authors defined a list of recommendation models considering recency, categories, and reading sequences among other factors. Their contextual bandit approach seeks to determine a strategy mapping models to contexts in order to maximize the expected rewards. They applied their contextual bandit both in NewsREEL Live and NewsREEL Replay. They report that performances vary depending on the domain under consideration.

### 3.3.2 The Algorithms in the Offline Task

The offline evaluation task has attracted several teams. The teams mainly focused on testing more sophisticated recommendation approaches (e.g. deep neural networks (Kumar et al. 2017)), studied efficient optimization of parameter configuration (e.g. finding similarity metrics for Collaborative Filtering (Beck et al. 2017)), and explored the technical complexity of algorithms. One advantage of the offline

task is that it does not require a permanent Internet connection and does not put additional burden on the participants to produce recommendations within a pre-defined tight time window. This ensured a low barrier to participate in the offline task and allowed participants to test new ideas and algorithms. In the remainder of this section, we discuss these, and other advantages, further.

*Ease of Use* Applying innovative ideas in a recommendation scenario typically requires extended testing and debugging. Before deploying algorithms, they are checked for their suitability to the scenario. The offline evaluation provides a well-suited environment for testing, debugging, and optimizing recommenders. Participants could simulate the stream on local hardware and study the strengths and weaknesses of new algorithms. The offline tests can control the load (by defining the number of concurrent messages sent by the offline simulation environment) and debug the functionality of the implemented solution. Participants typically tested algorithms first offline before moving to the online task. Innovative recommender approaches, for instance, based on Contextual Bandits or Deep Neural Networks have been evaluated offline.

*Parameter Optimization* Finding the optimal parameter configurations complements testing new approaches in offline evaluation. Optimization requires sufficiently large data streams to obtain robust results. Parallelization can be used to speed up optimization. The offline task supports parallelization. Participants can simulate the stream on multiple machines to arrive more quickly at the optimal configuration. In addition, the simulated stream can be replayed faster in order to accelerate the optimization process. The offline stream simulation ensures reproducible evaluation results as well as the comparability of the results obtained in different evaluation runs. This aspect of the offline task has been extensively used by several teams (e.g. by Beck et al. 2017).

*Technical Aspects* Tight time constraints, continuous changes of readers and articles, and the varying frequency with which messages emerge are difficult to simulate offline. The contest server allows participants to vary the number of concurrently sent messages. This facilitates finding bottlenecks which would cause errors in the online evaluation. Participants look at the distribution of response times to avoid such errors. This is particularly important for ensemble-based methods integration of multiple individual algorithms with varying complexities.

### 3.4 NewsREEL Evaluation

In order to evaluate the performance of different algorithms, NewsREEL followed the EaaS paradigm discussed by Hopfgartner et al. (2018).

In the four iterations of NewsREEL, most approaches achieved results superior to the baseline and still hold the potential for further optimization. The offline evaluation facilitates fine-grained analysis and parameter optimization for new

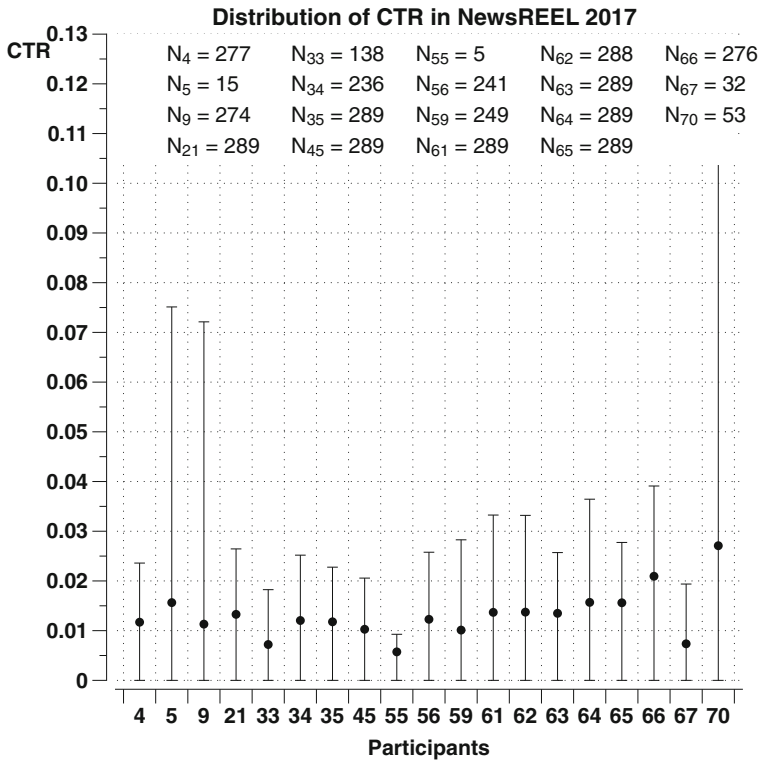


Fig. 4 Distribution of click-through rate in NewsREEL 2017

algorithms. Thereby, it enables participants to verify their ideas before deploying them online. The majority of participants used this opportunity. Figure 4 shows the distribution of click-through rate and standard deviation of all teams participating in NewsREEL 2017. In addition, the legend indicates for how many hours the corresponding algorithm had been active. A multitude of facets give rise to different perspectives on the quality desired of recommendations. First, we may ask who is to benefit from a recommender system? Readers, or users in general, avoid looking for information themselves. Publishers, on the other hand, retain readers and increase the chance for future visits. Second, we consider how to quantify utility. Recommendation has been modeled in various forms, including preference estimation, binary classification, and ranking problem. The click-through rate has been established as primary utility estimate in the online task. It represents the proportion of suggestions that readers clicked. Publishers would prefer to estimate their utility more directly, for instance, in terms of dwell time or the likelihood that readers will return. Both have proven difficult to compute with data available to NewsREEL. Sessions tend to include few reads which is why readers returning with the same session key are an uncommon phenomenon. In addition, computing

the dwell time requires the next read event. Moreover, a considerable subset of readers disallows session keys to be stored on their machines. As a result, we cannot distinguish them from one another rendering dwell time estimation impossible. Third, we have to take user experience into account. Waiting for recommendation entails costs similar to irrelevant recommendations. Readers are unlikely to wait for suggestions. Therefore, we have to consider additional aspects of utility such as availability, responsiveness, and scalability. In the online task, we monitor error events. These occur in cases when recommendation services fail to deliver in time or deliver invalid suggestions. In the offline task, the contest server computes the distribution of response times. This information enables us to compare algorithms in an additional dimension. i.e., it allows us to focus on both effectiveness and technical constraints that could not be evaluated in an online setting.

### 3.5 *NewsREEL Discussion*

The variety of methods used to address NewsREEL's tasks indicate a large number of connected research challenges for the future. While a more detailed analysis of these challenges is currently under preparation, we conclude this section by briefly highlighting the main successes and challenges of our initiative:

Successes:

- Being the first implementation of a living lab for the evaluation of information access systems, NewsREEL pioneered a new level of collaboration that enabled university-based researchers to gain access to a company's IT infrastructure and user base. We argue that this model of cooperation has the potential to narrow the growing gap between academic and commercial research in the field of information access.
- All of the four main information access evaluation campaigns (i.e., TREC, NTCIR, CLEF, and FIRE) have used news corpora in the past to advance research on challenges including ad-hoc retrieval, known item search, multilingual retrieval, and related retrieval tasks. NewsREEL contributes to this tradition by allowing further research on challenges such as real-time and stream processing, click optimization, and user profiling.
- NewsREEL has been used by practitioners, teachers at universities, and researchers. A survey amongst participants (Lommatzsch et al. 2017) has revealed that one of the main motivations for them to participate was to acquire new skills that are currently in high demand in industry. At the same time, NewsREEL has also been successfully embedded in teaching since students experienced factors associated with working in industry (Hopfgartner et al. 2016).



### Challenges:

- NewsREEL differs from the more traditional evaluation campaigns as participants had to ensure a high click-through rate under tight time constraints. We understand that these requirements were new to most researchers and that these different entry requirements might hold them back from participating. We addressed this by offering tutorials (e.g., at ECIR'15 (Hopfgartner and Brodt 2015) and ACM RecSys'15 (Hopfgartner et al. 2015b)), and by providing detailed instructions on how to get started on the NewsREEL website.
- In the online task, participants had to deal with fulfilling two goals at once. On the one hand, they had to optimize the click-through rate. On the other hand, they had to respond in a timely manner with valid items to guarantee a convenient user experience. The latter goal in particular, has caused major efforts as researchers tend to focus on algorithmic details rather than maintenance and scalability. Time constraints also had an effect on the computational complexity of algorithms. In addition, the real-time requirements render it difficult to debug the implementation. Although these are real issues and requirements that operators of online recommender systems face, we addressed this by introducing the offline task which allowed participants to implement and benchmark their algorithms and then deploy them to the online task.
- In the offline task, participants had to cope with the scale of the recorded data stream. Millions of events amount to gigabytes of data. Conducting experiments with the data takes a long time, in particular on personal computers. In order to address this, we released the benchmarking framework Idomaar that makes use of Big Data solutions such as Apache Kafka and Apache Flume. Idomaar can be deployed to Hadoop-based infrastructures that are able to cope with larger data streams (Scriminaci et al. 2016).
- In addition, the dynamic environment of news mandates continuous model updates. Seasonal trends, shifts in readers' interests, differences between working days and weekends or holidays produce varying behaviors of actors inside the news ecosphere. Breaking news events add another source for variation. This is in particular challenging for recommendation techniques that rely on exploiting users' prior interaction with news items (e.g., Hopfgartner and Jose 2014).
- The online component of NewsREEL causes additional challenges that need to be considered in order to guarantee a fair and unbiased evaluation. For example, some participants might suffer from network latency, especially if they were located far from plista's data centre in Germany. We addressed this limitation by offering virtual machines for participants in plista's data centre that they could use to deploy their algorithms. This solution is in line with the idea of EaaS as described by Hopfgartner et al. (2018).
- Receiving greatly varying numbers of requests can cause additional issues. For example, one participant may deliver a relatively high click-through rate with few requests, whereas another participants scores more clicks in total with more requests. Comparing these participants is difficult as the relatively high click-through rate could be due to chance.

## 4 Living Labs for Information Retrieval (LL4IR)

The main objective of the Living Labs for IR Evaluation (LL4IR) CLEF Lab was to provide a benchmarking platform for researchers to evaluate their ranking systems in a live setting. The lab acted as a proxy between commercial organizations (live environments) and lab participants (experimental systems), facilitated data exchange, and made comparisons between the participating systems. The lab focused on two use cases and one specific notion of what a living lab is. Use cases considered here were: product search (on an e-commerce site) and web search (through a commercial web search engine).

The LL4IR CLEF Lab contributed to the understanding of online evaluation as well as an understanding of the generalization of retrieval techniques across different use cases. Most importantly, it promoted IR evaluation that is more realistic, by allowing researchers to have access to historical search and usage data and by enabling them to validate their ideas in live settings with real users. This initiative was a first of its kind for IR.

This section reports on the results obtained during the official CLEF evaluation round that took place between May 1 and May 15, 2015. The positive feedback and growing interest from participants motivated us to organize a subsequent second unofficial evaluation round.

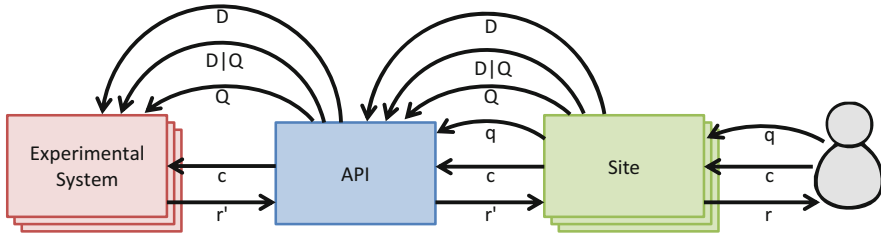
In the next section we describe the LL4IR API architecture and evaluation methodology. We then describe each of the two use cases in turn in Sects. 4.2 and 4.3, and provide details and analysis of the submissions received. In Sect. 4.4, we conclude with a discussion on LL4IR.

### 4.1 *LL4IR Architecture*

For the LL4IR CLEF Lab, evaluation was conducted primarily through an API. We first describe the workings of this API, followed by the evaluation setup divided into training and test phases. We then describe how we computed evaluation metrics using interleaved comparisons. Finally, we describe how we aggregated interleaving outcomes.

#### 4.1.1 LL4IR API

For each of the use cases, described in Sects. 4.2 and 4.3, challenge participants took part in a live evaluation process. For this they used a set of frequent queries as training queries and a separate set of frequent queries as test queries. Candidate documents were provided for each query and historical information associated with the queries. When participants produced their rankings for each query, they



**Fig. 5** Schematic representation of interaction with the LL4IR API, taken from Balog et al. (2014b)

uploaded these to the commercial provider use case through the provided LL4IR API. The commercial provider then interleaved a given participant’s ranked list with their own ranking, and presented the user with the interleaved result list. Participants took turns in having their ranked list interleaved with the commercial providers ranked list. This process of interleaving a single experimental system with the production system at a time was orchestrated by the LL4IR API, such that each participant gets about the same number of impressions. The actions performed by the commercial providers’ system users were then made available to the challenge participant (whose ranking had been shown) through the API; i.e., the interleaved ranking, resulting clicks, and (aggregated) interleaving outcomes.

Figure 5 shows the Living Labs architecture and how the participant interacted with the use cases through the LL4IR provided API. As can be seen, frequent queries ( $Q$ ) with candidate documents for each query ( $D|Q$ ) are sent from a site through the API to the experimental systems of participants. These systems upload their rankings ( $r'$ ) for each query to the API. When a user of the site issues one of these frequent queries ( $q$ ), then the site requests a ranking ( $r'$ ) from the API and presents it interleaved with  $r$  to the users. Any interactions ( $c$ ) of the user with this ranking are sent back to the API. Experimental systems can then obtain these interactions ( $c$ ) from the API and update their ranking ( $r'$ ) if they wish. We provided participants with example code and guidelines to ease the adaptation to our setup.<sup>3</sup> Our evaluation methodology, including reasons for focusing on frequent queries, is described in more detail in Balog et al. (2014b).

### 4.1.2 Training Phase

During the training phase, participants were free to update their rankings using feedback information. This feedback information was made available to them as soon as it arrived at the API. Their rankings could be updated at any time and as

<sup>3</sup><http://doc.living-labs.net/en/latest/guide-participant.html>.

often as desired. Both click feedback and aggregated outcomes were made available directly and were updated constantly.

### 4.1.3 Test Phase

In the test phase, challenge participants received another set of frequent queries as test queries. Again, the associated historical click information as well as candidate results for these queries were made available. After downloading the test queries, participants could only upload their rankings until the test phase started or only once after it started. These rankings were then treated in the same way as training queries. That is, they were interleaved with the commercial providers' rankings for several weeks. As for the training phase, in the test phase each challenge participant was given an approximately equal numbers of impressions. A major difference is that for the test queries, the click feedback is not made available. Aggregated outcomes were provided only after the test phase had ended.

### 4.1.4 Evaluation Metric

The overall evaluation of challenge participants was based on the final system performance, and additionally on how the systems performed at each query issue. The primary metric used was aggregated interleaving outcomes, and in particular the fractions of winning system comparisons. See Sect. 2.2 for details on interleaving comparisons. There are two reasons for using interleaved comparisons. Firstly, interleaved comparisons ensure that at least half the ranking shown to users comes from the production system. This reduces the risk of showing bad rankings to users. Secondly, interleaved comparisons were shown to be two orders of magnitude more sensitive than other ways of performing online evaluation such as A/B testing (Schuth et al. 2015c; Chapelle et al. 2012). As mentioned in Sect. 2.2, this means that far fewer query impressions are required to make informed decisions on which ranker gives better performance.

### 4.1.5 Aggregated Outcomes

LL4IR reported the following aggregated interleaving metrics, where *Outcome* served as the primary metric for comparing participants rankings. These aggregations were constantly updated for training queries. For the test phase they were only computed after the phase had finished.

*#Wins* is defined as the number of wins of the participant against the production system, where a *win* is defined as the experimental system having more clicks on results assigned to it by TDI than clicks on results assigned to the production system;

*#Losses* is defined as the number of losses against the production system;  
*#Ties* is defined as the number of ties with the production system;  
*#Impressions* is the total number of times when rankings (for any of the test queries) from the participant have been displayed to users of the production system; and  
*Outcome* is defined as the fraction of wins, so  $\#Wins / (\#Wins + \#Losses)$ .

An *Outcome* value below the *expected outcome* (typically 0.5) means that the participant system performed worse than the production system (i.e., overall it had more losses than wins). Significance of outcomes was tested using a two-sided binomial test which used the expected outcome and reported p-values.

Note that using these metrics, we are in theory only able to say something about the relationship between the participant's system and the production system. However, Radlinski et al. (2008) show experimentally that it is not unreasonable to assume transitivity. This allows us to also draw conclusions about how systems compare to each other. Ideally, instead of interleaving, we would have used multileaved comparison methods (Schuth et al. 2014, 2015b) which would directly give a ranking over rankers by comparing them all at once for each query.

## 4.2 LLAIR Use Case: Product Search

### 4.2.1 Task and Data

The *product search* use case is provided by REGIO Játék (REGIO Toy in English), the largest (offline) toy retailer in Hungary with currently over 30 stores. Their webshop<sup>4</sup> is among the top 5 in Hungary. The company is working on strengthening their online presence; improving the quality of product search in their online store is directed towards this larger goal. An excerpt from the search result page is shown in Fig. 6.

As described in Sect. 4.1, we distinguished training and test phases. Queries are sampled from the set of frequent queries; these queries are very short (1.18 terms on average) and have a stable search volume. For each query, a set of candidate products (approximately 50 products per query) and historical click information (click-through rate) was made available. For each product a structured representation was supplied (see below). The task then was to rank the provided candidate set.

---

<sup>4</sup><http://www.regiojatek.hu/>.

The screenshot shows the REGIO JATEK website interface. At the top, there is a search bar with the text 'angry birds' and a 'Keresés' button. Below the search bar, there are navigation tabs for 'KATEGÓRIÁK', 'ÉLETKOR', 'MÁRKÁK', 'MESEHŐS', 'AKCIÓK', and 'ÁRUHÁZAK'. A shopping cart icon labeled 'Kosár' with a '0' is visible in the top right. The main content area is titled 'Találatok' and shows 26 results. On the left, there are several filter sections: 'Kategóriák' (with checkboxes for Matrac, szőrf, rállós állatok, Készletfejlesztő, Papír, Irószerszám, Ősözgumi, karúszó, Akciófigurák), 'Márkák' (Hasbro, Bostway, Bestway), 'Mesehősök' (Angry Birds, Star Wars), and 'Nem' (mindegy). There is also an 'Életkor' slider and an 'Ár' slider. Below these filters, there are three columns of product listings. Each listing includes an image of the product, a title, and a price. The products shown are: 'Angry Birds - Star Wars kártya' (745 Ft), 'Angry Birds matricák ANG' (150 Ft), 'Angry Birds kártyagyűjtő album ANG' (695 Ft, 245 Ft), 'ANGRY BIRDS gyűjthető figurák, 2 db /cs' (2 130 Ft), 'Angry Birds SW. szivacsdobólgó 4 féléle A' (5 995 Ft), '2x90 db Angry Birds - Star Wars puzzle' (1 345 Ft, 745 Ft), 'Puzzle "4in1" Star Wars - Angry Birds' (2 255 Ft), 'Űszögumi Angry Birds 56cm' (745 Ft), and 'Angry Birds GO matrica ANG' (80 Ft). At the bottom left, there is a 'Hírlevél' section with a form for name and email, and a 'Feliratkozás' button.

Fig. 6 Screenshot of REGIO, the LLAIR product search use case

**Table 1** Fielded document representation of products in the LL4IR product search use case

| Field             | Description  |
|-------------------|--|
| age_max           | Recommended maximum age (may be empty, i.e., 0)  |
| age_min           | Recommended minimum age (may be empty, i.e., 0)  |
| arrived           | When the product arrived (first became available); only for products that arrived after 2014-08-28 |
| available         | Indicates if the product is currently available (1) or not (0)                                     |
| bonus_price       | Provided only if the product is on sale; this is the new (sales) price                             |
| brand             | Name of the brand (may be empty)   |
| category          | Name of the (leaf-level) product category  |
| category_id       | Unique ID of the (leaf-level) product category   |
| characters        | List of toy characters associated with the product (may be empty)                                  |
| description       | Full textual description of the product (may be empty)   |
| main_category     | Name of the main (top-level) product category  |
| main_category_id  | Unique ID of the main (top-level) product category   |
| gender            | Gender recommendation. (0: for both girls and boys (or unclassified); 1: for boys; 2: for girls)   |
| photos            | List of photos about the product   |
| price             | Normal price   |
| product_name      | Name of the product  |
| queries           | Distribution of (frequent) queries that led to this product (may be empty)                         |
| short_description | Short textual description of the product (may be empty)  |

### 4.2.2 Product Descriptions

For each product a fielded document representation was provided, containing the attributes shown in Table 1. The amount of text available for individual products is limited (and is in Hungarian), but there are structural and semantic annotations, including:

- Organization of products into a two-level deep topical categorization system;
- Toy characters associated with the product (Barbie, Spiderman, Hello Kitty, etc.);
- Brand (Beados, LEGO, Simba, etc.);
- Gender and age recommendations (for many products);
- Queries (and their distribution) that led to the given product.

### 4.2.3 Candidate Products

The candidate set, to be ranked, contained all products that were available in the (recent) past. This comprises all products that were considered by the site’s production search engine (in practice: all products that contain any of the query terms in any of their textual fields). One particular challenge for this use case is that the inventory (as well as the prices) are constantly changing; however, for

challenge participants, a single ranking is used throughout the entire test period of the challenge, without the possibility of updating it. The candidate set therefore also includes products that may not be available at the moment (but might become available again in the future). Participating systems were strongly encouraged to consider all products from the provided candidate set. Those that were unavailable at a given point in time were not displayed to users of the REGIO online store. Further, it might happen (and as we show in Schuth et al. (2015a) it indeed did happen) during the test period that new products arrive; experimental systems were unable to include these in their ranking (this was the same for all participants), while the production system might return them. This can potentially affect the number of wins against the production system (to the advantage of the production system), but it does not affect the comparison across experimental systems.

#### 4.2.4 Submissions and Results

Two organizations submitted a total of four runs. In addition, a simple baseline provided by the challenge organizers was also included for reference. Table 2 presents the results.

#### 4.2.5 Approaches

The organizers' baseline (BASELINE in Table 2) ranks products based on historical click-through rate. Only products that were clicked for the given query are returned; their attributes are ignored. In case historical clicks are unavailable (this happened for a single query R- $\alpha$ 97), (all) candidate products are returned in an arbitrary order (in practice, in the same order as they were received from the API via the `doclist` request).

The University of Stavanger (Ghirmatsion and Balog 2015) employed a fielded document retrieval approach based on language modeling techniques. Specifically, building upon the Probabilistic Retrieval Model for Semistructured Data by Kim et al. (2009), they experimented with three different methods (UIS-\*) for estimating term-field mapping probabilities. Their results show that term-specific field mapping

**Table 2** Results for the product search use case

| Submission                               | Outcome | #Wins | #Losses | #Ties | #Impressions | <i>p</i> -value |
|--|---------|-------|---------|-------|--------------|-----------------|
| BASELINE                                 | 0.4691  | 91    | 103     | 467   | 661          | <0.01           |
| UIS-MIRA (Ghirmatsion and Balog 2015)    | 0.3413  | 71    | 137     | 517   | 725          | 0.053           |
| UIS-JERN (Ghirmatsion and Balog 2015)    | 0.3277  | 58    | 119     | 488   | 665          | 0.156           |
| UIS-UIS (Ghirmatsion and Balog 2015)     | 0.2827  | 54    | 137     | 508   | 699          | 0.936           |
| GESIS (Schaer and Tavakolpoursaleh 2015) | 0.2685  | 40    | 109     | 374   | 523          | 0.785           |

The expected outcome under a randomly clicking user is 0.28. *P*-values are computed using a binomial test



in general is beneficial, but their attempt at estimating field importance based on historical click-through information met with limited success.

Team GESIS (Schaer and Tavakolpoursaleh 2015) also used a fielded document representation. They used Solr for ranking products and incorporated historical click-through rates, if available, as a weighting factor.

#### 4.2.6 Dealing with Inventory Changes

As mentioned in Sect. 4.2.1, the product inventory is subject to changes. Not all products that were part of the candidate set were available at all times. If all products were available, the expected probability of winning an interleaved comparison (assuming a randomly clicking user) would be 0.5. However, on average, 44% of the products were actually unavailable. These products were only ever present in the participants' ranking (the site's ranking never considered them). And, only *after* interleaving were these products removed from the resulting interleaved list. We note that this is undesired behavior, as they should have been filtered out *before* interleaving. The necessary adjustments were made to the implementation for the next round of the challenge. As for interpreting these results, this means that the chances for products from the participants ranking to be clicked were reduced. This in turn reduced the expected probability to win to:

$$\Pr(\text{participant} > \text{site}) = (1 - 0.44) \cdot 0.5 = 0.28.$$

Consequently, if a participant's system wins more than in 28% of the impressions, then this is more than expected. And thus the participant's system can be said to be better than the site's system if the outcome is (significantly) more than 28%.

#### 4.2.7 Results

We find that at least three submissions are likely to have improved upon the production system's ranking. Somewhat surprisingly, the simple baseline performed by far the best, with an outcome of 0.4691. This was also the only system that significantly outperformed the production system. The best performing participant run is UIS-MIRA, with an outcome of 0.3413. A more in-depth analysis of the results is provided in the LL4IR extended lab overview paper (Schuth et al. 2015a).

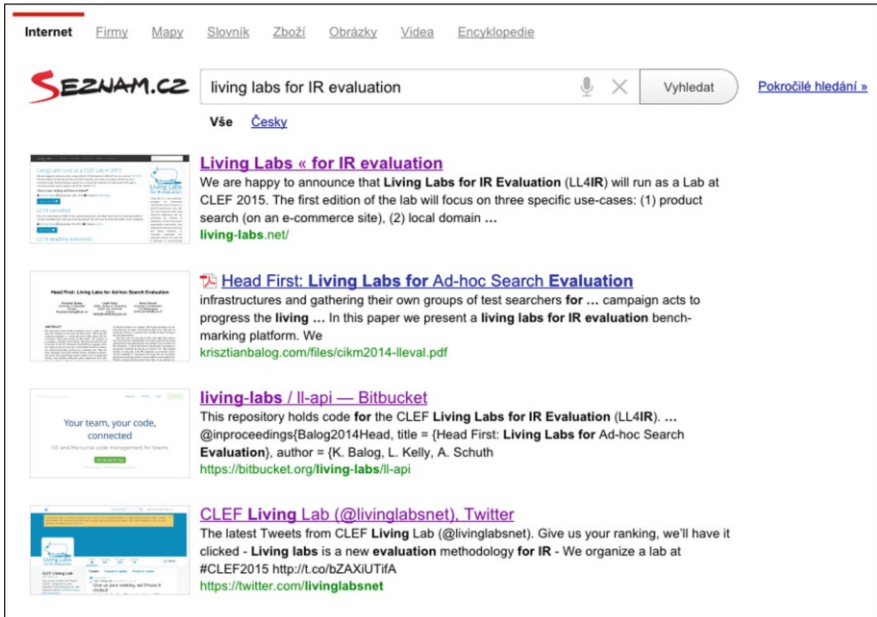


Fig. 7 Screenshot of Seznam, the LL4IR web search use case

### 4.3 LL4IR Use Case: Web Search

#### 4.3.1 Task and Data

The *web search* use case has been provided by Seznam,<sup>5</sup> a very large web search engine in the Czech Republic. See Fig. 7 for a screenshot of the user interface.

Seznam serves almost half the country's search traffic and as such has very high site traffic. Queries are the typical web search queries, and thus are a mixed bag of navigational and transactional (Broder 2002). In contrast to the product search use case, apart from the scale and the query types, Seznam did not make raw document and query content available, rather features computed for documents and queries. This is much like any learning to rank dataset, such as Letor (Liu et al. 2007). Queries and documents are only identified by a unique identifier and for each query, the candidate documents are represented with sparse feature vectors. Seznam provided a total of 557 features. These features were not described in any way. The challenge with this use case then is a learning to rank challenge (Liu 2009).

As described in Sect. 4.1, the web search use case also consists of a training and test phase. For the test phase, there were 97 queries, for the training phase 100

<sup>5</sup><http://search.seznam.cz/>.

**Table 3** Results for the web search use case

| Submission            | Outcome | #Wins | #Losses | #Ties  | #Impressions | <i>p</i> -value |
|-----------------------|---------|-------|---------|--------|--------------|-----------------|
| EXPLOITATIVE BASELINE | 0.5527  | 3030  | 2452    | 19,055 | 24,537       | <0.01           |
| UNIFORM BASELINE      | 0.2161  | 430   | 1560    | 1346   | 3336         | <0.01           |

The expected outcome under a randomly clicking user is 0.5. *P*-values were computed using a binomial test

queries were provided. On average, for each query there were about 179 candidate documents. In total, there were 35,322 documents.

### 4.3.2 Results

The web search use case attracted six teams that submitted runs for the training queries. However, none of them submitted runs for the test queries. Therefore, we can only report on two baseline systems, provided by the challenge organizers. Baseline 1, titled EXPLOITATIVE BASELINE in Table 3, uses the original Seznam ranking and was therefore expected to produce an outcome of 0.5.<sup>6</sup> Baseline 2, titled UNIFORM BASELINE in Table 3, assigned uniform weights to each feature and ranked by the weighted sum of feature values. This baseline was expected not to perform well.

There were over 440K impressions on Seznam through our Living Labs API. On average this amounts to 2247 impressions for each query. Approximately 6% of all impressions were used for the testing period. As can be seen in Table 3, the EXPLOITATIVE BASELINE outperformed the production system. An outcome (outcome measure described in Sect. 4.1) of 0.5527 has been achieved, with 3030 wins and 2452 losses against the production system, and 19,055 ties with it. As expected, the UNIFORM BASELINE lost many more comparisons than it won. Both outcomes were statistically significant according to a binomial test. Again, we refer to the LL4IR extended lab overview paper (Schuth et al. 2015a) for full details.

## 4.4 LL4IR Discussion

The living labs methodology offers great potential to evaluate information retrieval systems in live settings with real users. The LL4IR CLEF Lab represents the first attempt at a shared community benchmarking platform in this space. The first edition of LL4IR focused on two use-cases, product search and web search, using a

<sup>6</sup>If use cases uploaded their candidate documents in the order that represented their own ranking, then this was available to participants. We plan to change this in the future.

commercial e-commerce website, REGIO, and a commercial web search engine, Seznam. Below, we identify some of the main successes and challenges of our initiative.

Successes:

- A major contribution of the lab is the development of the necessary API infrastructure, which has been made publicly available. Overall, we regard our effort successful in showing the feasibility and potential of this form of evaluation. For both use-cases, there was an experimental system that outperformed the corresponding production system significantly. It is somewhat unfortunate that in both cases that experimental system was a baseline approach provided by the challenge organizers, nevertheless, it demonstrates the potential benefits to use-case owners as well.
- The API infrastructure developed for the LL4IR CLEF Lab offers the potential to host ongoing IR evaluations in a live setting. As such, it is planned that these “challenges” will continue on an ongoing basis post-CLEF, with an expanding number of use-cases as well as refinements to the existing use-cases.<sup>7</sup> A more detailed analysis of the use-cases, including results from a second unofficial evaluation round, and a discussion of ideas and opportunities for future development is provided in the LL4IR extended lab overview paper (Schuth et al. 2015a).

Challenges:

- *Startup challenge:* The LL4IR CLEF Lab attracted interest from dozens of teams. There were twelve active participants, but only two teams ended up submitting results for the official evaluation (excluding the organizers’ baseline systems). We found that, while many researchers expressed and showed their interest in the lab, our setup with an API, instead of a static test collection, was a hurdle for many. We plan to ease this process of adapting to this new evaluation paradigm by providing even more examples and by organizing tutorials where we demonstrate working with our API.
- *Frequency of inventory change:* One particular issue that surfaced and needs addressing for the product search use-case is the frequent changes in inventory. This appears to be more severe than we first anticipated and represents some challenges, both technical and methodological.

## 5 Discussion and Conclusion

In this chapter, we have discussed the importance of conducting online evaluations using real participants conducting real tasks in the wild. We have presented two evaluation initiatives which address this need by offering shared challenges which

---

<sup>7</sup>See <http://living-labs.net/> for details.

**Table 4** Comparison of static test collections and living labs

|                       | Static test collections  | Living labs   |
|-----------------------|--|---|
| Representativeness    | Data is only as good as the guidelines                                 | Real user data, real and representative information needs                                       |
| Scalability           | Not scalable in terms of users; very scalable in terms of participants | Very scalable in terms of users; for participants, scalability is limited by the site's traffic |
| Effort (organizers)   | One-off  | Continuous  |
| Effort (participants) | Moderate   | Increased   |
| Reproducibility       | Results of previous approaches are easily reproducible                 | For a fair comparison, a new online evaluation round is needed                                  |

operate in a living labs setting. Specifically, the NewsREEL shared challenge for recommender systems, and the LL4IR shared challenge for information retrieval. The aim of these initiatives is to close the gap that exists between industry and academia in the evaluation of information access systems. Both campaigns can be seen as initiatives that follow the Evaluation-as-a-Service paradigm discussed by Hopfgartner et al. (2018).

We argue that access to living labs style shared challenges, which offer researchers the opportunity to evaluate their algorithms in an online setting with real users of systems, is essential for researchers to be able to study the performance of algorithms under real-world conditions. However, although continuous evaluation of large-scale information access systems is clearly an important tool for advancing the state of the art, we cannot expect living labs to arise spontaneously and automatically. Instead, creating and running initiatives that offer online opportunities for evaluation requires the investment of resources and a great deal of persistence on the part of organizers and participants. A detailed discussion on key technical aspects and efforts required to establish Evaluation-as-a-Service as a mature evaluation methodology is provided by Hopfgartner et al. (2018). Extending on their discussion, we close this chapter by highlighting reasons that illustrate the necessity to continue to invest effort into promoting the living labs online evaluation paradigm. As summarized in Table 4, we concentrate our discussion on the differences between traditional evaluation campaigns based on static datasets and living lab campaigns.

- *Representativeness*: As discussed earlier, static test collections have played a significant role in the evaluation of information access methods. In fact, for many years, test collections and related shared evaluation tasks were used to define and to study current research challenges. In the past few years, however, we could observe a paradigm shift, where commercial research on information access systems relies increasingly on online benchmarking, also referred to as A/B testing. The reason for this development is that users and their information needs have become a significant factor that affects retrieval and recommendation algorithms. Static test collections, however, are often not suitable for the development of user-centric techniques. First of all, the need

to define search tasks might not really reflect users' real information needs. In addition, relevance judgements might be highly subjective and therefore could have a negative effect on personalization techniques. In addition, the dataset used might not be suitable, e.g., because it is outdated or because the users are not interested in its content. Living labs as described in this chapter, however, can help us to reduce these negative effects. They enable us to rely on real user interactions, i.e., users use the living lab service to satisfy their personal information needs. This allows us to avoid negative factors such as the observer expectancy effect that could impact any type of personalization method.

- *Scalability*: For many years, interactive information access methods were evaluated in relatively small user experiments with a limited number of search tasks and participants. For a detailed discussion on this, we refer the reader to Sakai (2018). University-based researchers in particular employed these small-scale experiments since they often lack access to resources required to perform larger user studies. Industry-based researchers, however, often have access to a large number of users and consequently, large-scale user experiments can nowadays be seen as the de-facto evaluation standard. This differing access to resources, however, has led to a growing gap between academia and industry. Living labs can help in narrowing this gap since they can enable university-based researchers to gain access to a larger user base.
- *Effort (Organizers)*: One of the main advantages of shared evaluation tasks is that the effort that goes into their organization is restricted. Although work involved such as defining tasks, document procurement, topic development, conducting experiments, developing relevance assessments, or evaluating results can be time consuming, they only have to be performed once. Living labs, however, require a continuous efforts from the organizers since they have to guarantee that the live service as well as all technical components that are involved in the evaluation campaign remain fully functional.
- *Effort (Participants)*: One of the main advantages of shared evaluation campaigns that rely on static data collections is that these campaigns are often organized in a very similar fashion. Usually, participants are required to produce a ranked list of retrieval results for a given dataset and search task. Then, standard evaluation metrics are calculated, e.g., using the popular tool `trec_eval`.<sup>8</sup> Given this “standardized” approach, experienced information access researchers might find it easier to participate in these tasks since they have to put less effort into understanding the evaluation process. Living labs, however, are more demanding. For example, in NewsREEL, participants need to set up their own server and register it with the open recommendation platform to gain access to the data. Further, they have to make sure that their system is running smoothly over a longer period of time. Our observation from running NewsREEL is that implementing stable solutions that are able to operate over a longer time period was challenging for many participants.

---

<sup>8</sup>[https://github.com/usnistgov/trec\\_eval](https://github.com/usnistgov/trec_eval).

- *Reproducibility*: Scientific progress requires accumulating experimental findings that are reproducible, i.e., ensuring that the findings of testing an algorithm on a test collection can be recreated by another team, thus enabling the new team to develop new approaches and compare them to the first approach. Freire et al. (2016) discuss challenges related to reproducibility in offline data-oriented experiments in detail. The authors point out that reproducibility is made difficult by *volatility of the data*, pointing to the example of live streams in which the same situation never occurs again. Future work is needed in order to set up guidelines for reproducing an experiment without using exactly the same data. A related question is the ability to predict the results of online evaluation using offline experiments. We remark that most discussions on reproducibility assume that the evaluation metric is fixed. However, for information access systems, the ideal goal is to ensure that research results can be reproduced in terms of success criteria that go beyond specific evaluation metrics. User satisfaction is a key success criterion, yet, success has many facets (see, e.g., multi-dimensional evaluation models for recommender systems (Said et al. 2012)). It is clear that further work is needed on the development of metrics for evaluating the success of information access systems. Such work will help to further develop the usefulness of both the offline and the online evaluation paradigms.

In summary, there appears to be general agreement that the future of the evaluation of information access systems lies in evaluating under ever-more realistic conditions. In this chapter, we have emphasized the necessity for public benchmarks offering the possibility to test information access systems online in order to bridge the gap between academia and industry. Here, we would also like to point out that industry also stands to benefit from online evaluation initiatives. Internally, a company can only test their own algorithms on their own data stream. Online evaluations offer a valuable opportunity to test algorithms head-to-head with the full range of participating algorithms on other data streams. The widespread agreement on the value of online evaluation stands in contrast to the relatively slow pace at which online evaluation has begun to be adopted in the research community. Our hope is that the motivation and description of online evaluation provided in this chapter will encourage others to continue to invest effort in evaluation that will allow continuous evaluation of large-scale information access systems to realize its full potential.

## References

- Allan J, Croft B, Moffat A, Sanderson M (2012) Frontiers, challenges, and opportunities for information retrieval: report from SWIRL 2012 the second strategic workshop on information retrieval in Lorne. SIGIR Forum 46(1):2–32
- Azzopardi L, Balog K (2011) Towards a living lab for information retrieval research and development - a proposal for a living lab for product search tasks. In: Forner P, Gonzalo J, Kekäläinen J, Lalmas M, de Rijke M (eds) Multilingual and multimodal information

- access evaluation. Proceedings of the second international conference of the cross-language evaluation forum (CLEF 2011). Lecture notes in computer science (LNCS), vol 6941. Springer, Heidelberg, pp 26–37
- Balog K, Elsweiler D, Kanoulas E, Kelly L, Smucker MD (2014a) Report on the CIKM workshop on living labs for information retrieval evaluation. SIGIR Forum 48(1):21–28
- Balog K, Kelly L, Schuth A (2014b) Head first: living labs for ad-hoc search evaluation. In: Proceedings of the 23rd international conference on information and knowledge management (CIKM'14). ACM, New York, pp 1815–1818
- Beck PD, Blaser M, Michalke A, Lommatzsch A (2017) A system for online news recommendations in real-time with Apache mahout. In: Working notes of the 8th international conference of the CLEF initiative, Dublin, CEUR workshop proceedings
- Bons P, Evans N, Kampstra P, van Kessel T (2017) A news recommender engine with a killer sequence. In: Working notes of the 8th international conference of the CLEF initiative, Dublin, CEUR workshop proceedings
- Broder A (2002) A taxonomy of web search. SIGIR Forum 36(2):3–10
- Brodth T, Hopfgartner F (2014) Shedding light on a living lab: the CLEF NewsREEL open recommendation platform. In: Proceedings of the information interaction in context conference, IiiX'14. Springer, New York, pp 223–226
- Chapelle O, Joachims T, Radlinski F, Yue Y (2012) Large-scale validation and analysis of interleaved search evaluation. ACM Trans Info Syst (TOIS) 30:1–41
- Ciobanu A, Lommatzsch A (2016) Development of a news recommender system based on Apache flink. In: Working notes of the 7th international conference of the CLEF initiative, Evora, CEUR workshop proceedings
- Corsini F, Larson M (2016) CLEF NewsREEL 2016: image based recommendation. In: Working notes of the 7th international conference of the CLEF initiative, Evora, CEUR workshop proceedings
- Diaz F, White R, Buscher G, Liebling D (2013) Robust models of mouse movement on dynamic web search results pages. In: Proceedings of the 22nd ACM international conference on information and knowledge management (CIKM'13), pp 1451–1460
- Domann J, Meiners J, Helmers L, Lommatzsch A (2016) Real-time news recommendations using Apache spark. In: Working notes of the 7th international conference of the CLEF initiative, Evora, CEUR workshop proceedings
- Freire J, Fuhr N, Rauber A (2016) Reproducibility of data-oriented experiments in e-science (Dagstuhl seminar 16041). In: Dagstuhl reports, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, vol 6
- Gebremeskel G, de Vries AP (2015) The degree of randomness in a live recommender systems evaluation. In: Working notes for CLEF 2015 conference, Toulouse, CEUR
- Ghirmatsion AB, Balog K (2015) Probabilistic field mapping for product search. In: CLEF 2015 online working notes
- Golian C, Kuchar J (2017) News recommender system based on association rules at CLEF NewsREEL 2017. In: Working notes of the 8th international conference of the CLEF initiative, Dublin, CEUR workshop proceedings
- Gunawardana A, Shani G (2009) A survey of accuracy evaluation metrics of recommendation tasks. J Mach Learn Res 10:2935–2962
- Hanbury A, Müller H, Balog K, Brodt T, Cormack GV, Eggel I, Gollub T, Hopfgartner F, Kalpathy-Cramer J, Kando N, Krithara A, Lin JJ, Mercer S, Potthast M (2015) Evaluation-as-a-service: overview and outlook. CoRR abs/1512.07454
- Hassan A, Shi X, Craswell N, Ramsey B (2013) Beyond clicks: query reformulation as a predictor of search satisfaction. In: Proceedings of the 22nd ACM international conference on information and knowledge management (CIKM'13). ACM, New York, pp 2019–2028
- Hawking D (2015) If SIGIR had an academic track, what would be in it? In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, Santiago, August 9–13, 2015, p 1077



- Hofmann K, Whiteson S, de Rijke M (2011) A probabilistic method for inferring preferences from clicks. In: Proceedings of the 20th conference on information and knowledge management (CIKM'11). ACM, New York, p 249
- Hopfgartner F, Brodt T (2015) Join the living lab: evaluating news recommendations in real-time. In: Advances in information retrieval - 37th European conference on IR research, ECIR 2015, Proceedings, Vienna, March 29–April 2, 2015, pp 826–829
- Hopfgartner F, Jose JM (2014) An experimental evaluation of ontology-based user profiles. *Multimed Tools Appl* 73(2):1029–1051
- Hopfgartner F, Kille B, Lommatzsch A, Plumbaum T, Brodt T, Heintz T (2014) Benchmarking news recommendations in a living lab. In: Kanoulas E, Lupu M, Clough P, Sanderson M, Hall M, Hanbury A, Toms E (eds) Information access evaluation – multilinguality, multimodality, and interaction. Proceedings of the fifth international conference of the CLEF initiative (CLEF 2014). Lecture notes in computer science (LNCS), vol 8685. Springer, Heidelberg, pp 250–267
- Hopfgartner F, Brodt T, Seiler J, Kille B, Lommatzsch A, Larson M, Turrin R, Serény A (2015a) Benchmarking news recommendations: the CLEF newsreel use case. *SIGIR Forum* 49(2):129–136
- Hopfgartner F, Kille B, Heintz T, Turrin R (2015b) Real-time recommendation of streamed data. In: Proceedings of the 9th ACM conference on recommender systems, RecSys 2015, Vienna, September 16–20, 2015, pp 361–362
- Hopfgartner F, Lommatzsch A, Kille B, Larson M, Brodt T, Cremonesi P, Karatzoglou A (2016) The potentials of recommender systems challenges for student learning. In: Proceedings of CiML'16: challenges in machine learning: gaming and education
- Hopfgartner F, Hanbury A, Mueller H, Eggel I, Balog K, Brodt T, Cormack GV, Lin J, Kalpathy-Cramer J, Kando N, Kato MP, Krithara A, Gollub T, Potthast M, Viegas E, Mercer S (2018) Evaluation-as-a-service for the computational sciences: overview and outlook. *ACM J Data Inf Qual*. <https://doi.org/10.1145/3239570>
- Jagerman R, Balog K, de Rijke M (2018) Opensearch: lessons learned from an online evaluation campaign. *J Data Inf Qual* 10(3):13:1–13:15
- Joachims T (2003) Evaluating retrieval performance using clickthrough data. In: Franke J, Nakhaeizadeh G, Renz I (eds) Text mining. Physica. Springer, Heidelberg, pp 79–96
- Joachims T, Granka LA, Pan B, Hembrooke H, Radlinski F, Gay G (2007) Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans Inf Syst* 25(2):7
- Kamps J, Geva S, Peters C, Sakai T, Trotman A, Voorhees E (2009) Report on the SIGIR 2009 workshop on the future of IR evaluation. *SIGIR Forum* 43(2):13–23
- Kelly D, Dumais ST, Pedersen JO (2009) Evaluation challenges and directions for information-seeking support systems. *IEEE Comput* 42(3):60–66
- Kelly L, Bunbury P, Jones GJF (2012) Evaluating personal information retrieval. In: Proceedings of the 34th European conference on information retrieval (ECIR'12). Springer, Berlin
- Kille B, Hopfgartner F, Brodt T, Heintz T (2013) The plista dataset. In: NRS'13: proceedings of the international workshop and challenge on news recommender systems. ACM, New York, pp 14–21
- Kille B, Lommatzsch A, Turrin R, Serény A, Larson M, Brodt T, Seiler J, Hopfgartner F (2015) Stream-based recommendations: online and offline evaluation as a service. In: Mothe J, Savoy J, Kamps J, Pinel-Sauvagnat K, Jones GJF, SanJuan E, Cappellato L, Ferro N (eds) Experimental IR meets multilinguality, multimodality, and interaction. Proceedings of the sixth international conference of the CLEF association (CLEF 2015). Lecture notes in computer science (LNCS), vol 9283. Springer, Heidelberg, pp 497–517
- Kim J, Xue X, Croft WB (2009) A probabilistic retrieval model for semistructured data. In: Proc. of the 31st European conference on information retrieval (ECIR'09). Springer, Heidelberg, pp 228–239
- Kim Y, Hassan A, White R, Zitouni I (2014) Modeling dwell time to predict click-level satisfaction. In: Proc. of the 7th ACM international conference on web search and data mining (WSDM'14). ACM, New York, pp 193–202

- Kohavi R (2015) Online controlled experiments: lessons from running A/B/n tests for 12 Years. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, Sydney, NSW, August 10–13, 2015, p 1
- Kumar V, Khattar D, Gupta S, Gupta M, Varma V (2017) Deep neural architecture for news recommendation. In: Working notes of the 8th international conference of the CLEF initiative, Dublin, CEUR workshop proceedings
- Li J, Huffman S, Tokuda A (2009) Good abandonment in mobile and pc internet search. In: Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval (SIGIR '09). ACM, New York, pp 43–50
- Liang Y, Loni B, Larson M (2017) CLEF NewsREEL 2017: contextual bandit news recommendation. In: Working notes of the 8th international conference of the CLEF initiative, Dublin, CEUR workshop proceedings
- Liu TY (2009) Learning to rank for information retrieval. *Found Trends Inf Retr* 3(3):225–331
- Liu TY, Xu J, Qin T, Xiong W, Li H (2007) LETOR: benchmark dataset for research on learning to rank for information retrieval. In: Proceedings of SIGIR 2007 workshop on learning to rank for information retrieval (LR4IR'07), pp 346–374
- Lommatzsch A, Albayrak S (2015) Real-time recommendations for user-item streams. In: Proc. of the 30th symposium on applied computing, SAC 2015, SAC '15. ACM, New York, pp 1039–1046
- Lommatzsch A, Johannes N, Meiners J, Helmers L, Domann J (2016) Recommender ensembles for news articles based on most-popular strategies. In: Working notes of the 7th international conference of the CLEF initiative, Evora, CEUR workshop proceedings
- Lommatzsch A, Kille B, Hopfgartner F, Larson M, Brodt T, Seiler J, Özgöbek Ö (2017) CLEF 2017 NewsREEL overview: a stream-based recommender task for evaluation and education. In: Jones GJF, Lawless S, Gonzalo J, Kelly L, Goeuriot L, Mandl T, Cappellato L, Ferro N (eds) *Experimental IR meets multilinguality, multimodality, and interaction*. Proceedings of the eighth international conference of the CLEF association (CLEF 2017). Lecture notes in computer science (LNCS), vol 10456. Springer, Heidelberg, pp 239–254
- Ludmann C (2017) Recommending news articles in the CLEF news recommendation evaluation lab with the data stream management system odysseus. In: Working notes of the 8th international conference of the CLEF initiative, Dublin, CEUR workshop proceedings
- Radlinski F, Craswell N (2013) Optimized interleaving for online retrieval evaluation. In: Proc. of ACM international conference on web search and data mining (WSDM'13). ACM, New York, pp 245–254
- Radlinski F, Kurup M, Joachims T (2008) How does clickthrough data reflect retrieval quality? In: Proceedings of the 17th conference on information and knowledge management (CIKM'08). ACM, New York, pp 43–52
- Said A, Tikk D, Stumpf K, Shi Y, Larson M, Cremonesi P (2012) Recommender systems evaluation: a 3D benchmark. In: Proceedings of the workshop on recommendation utility evaluation: beyond RMSE (RUE 2012), CEUR-WS, vol 910, RUE'12, pp 21–23
- Sakai T (2018) *Laboratory experiments in information retrieval*. Springer, Singapore
- Schaer P, Tavakolpoursaleh N (2015) GESIS at CLEF LL4IR 2015. In: CLEF 2015 Online Working Notes
- Schuth A, Sietsma F, Whiteson S, Lefortier D, de Rijke M (2014) Multileaved comparisons for fast online evaluation. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM'14). ACM, New York, pp 71–80
- Schuth A, Balog K, Kelly L (2015a) Extended overview of the living labs for information retrieval evaluation (LL4IR) CLEF lab 2015. In: CLEF 2015 online working notes
- Schuth A, Bruintjes RJ, Büttner F, van Doorn J, Groenland C, Oosterhuis H, Tran CN, Veeling B, van der Velde J, Wechsler R, Woudenberg D, de Rijke M (2015b) Probabilistic multileave for online retrieval evaluation. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval (SIGIR'15). ACM, New York, pp 955–958

- Schuth A, Hofmann K, Radlinski F (2015c) Predicting search satisfaction metrics with interleaved comparisons. In: Proceedings of the 38th ACM international conference on information retrieval (SIGIR'15). ACM, New York pp 463–472
- Scriminaci M, Lommatzsch A, Kille B, Hopfgartner F, Larson M, Malagoli D, Serény A, Plumbaum T (2016) Idomaar: a framework for multi-dimensional benchmarking of recommender algorithms. In: Proceedings of the poster track of the 10th ACM conference on recommender systems (RecSys 2016), Boston, September 17, 2016
- Tavakolifard M, Gulla JA, Almeroth KC, Hopfgartner F, Kille B, Plumbaum T, Lommatzsch A, Brodt T, Bucko A, Heintz T (2013) Workshop and challenge on news recommender systems. In: Seventh ACM conference on recommender systems, RecSys '13, Hong Kong, October 12–16, 2013, pp 481–482
- Teevan J, Dumais S, Horvitz E (2007) The potential value of personalizing search. In: Proceedings of the ACM international conference on information retrieval (SIGIR'07). ACM, New York, pp 756–757
- Turpin A, Scholar F (2006) User performance versus precision measures for simple search tasks. In: Proc. of the ACM international conference on information retrieval (SIGIR'06). ACM, New York, pp 11–18
- Verbitskiy I, Probst P, Lommatzsch A (2015) Developing and evaluation of a highly scalable news recommender system. In: Working notes for CLEF 2015 conference, Toulouse, CEUR
- Voorhees EM, Harman DK (2005) TREC: Experiment and evaluation in information retrieval, 1st edn. MIT Press, Cambridge, MA
- Wang K, Gloy N, Li X (2010) Inferring search behaviors using partially observable Markov (POM) model. In: WSDM'10. ACM, New York, pp 211–220
- Wilkins P, Byrne D, Jones GJF, Lee H, Keenan G, McGuinness K, O'Connor NE, O'Hare N, Smeaton AF, Adamek T, Troncy R, Amin A, Benmokhtar R, Dumont E, Huet B, Mérialdo B, Toliaas G, Spyrou E, Avrithis YS, Papadopoulos GT, Mezaris V, Kompatsiaris I, Mörzinger R, Schallauer P, Bailer W, Chandramouli K, Izquierdo E, Goldmann L, Haller M, Samour A, Cobet A, Sikora T, Praks P, Hannah D, Halvey M, Hopfgartner F, Villa R, Punitha P, Goyal A, Jose JM (2008) K-space at TRECVID 2008. In: TRECVID 2008 workshop participants notebook papers, Gaithersburg, MD, Nov 2008
- Yilmaz E, Verma M, Craswell N, Radlinski F, Bailey P (2014) Relevance and effort: an analysis of document utility. In: Proceedings of the 23rd ACM international conference on information and knowledge management (CIKM'14). ACM, New York, pp 91–100

**Part VI**  
**Impact and Future Challenges**

# The Scholarly Impact of CLEF 2010–2017



## A Google Scholar Analysis of CLEF Proceedings and Working Notes

**Birger Larsen**

**Abstract** This chapter assesses the scholarly impact of the CLEF evaluation campaign by performing a bibliometric analysis of the citations of the CLEF 2010–2017 papers collected through Google Scholar. The analysis extends an earlier 2013 study by Tsikrika et al. of the CLEF Proceedings for the period 2000–2009 and compares the impact of the first half of CLEF to the second. It also extends the analysis by including the CLEF Working notes, a less formal but important part of the CLEF oeuvre. Results show that, despite the different nature of the peer-reviewed CLEF Proceedings papers and the less formal and much more numerous Working note papers, both types of publications have high citation impact. In particular, overview papers from the various labs and tasks in CLEF attract large amounts of citations in both Proceedings and Working Notes. A significant proportion of the total number of citations appear to be from outside CLEF—there are simply not enough CLEF papers every year to explain that many citations. In conclusion, the analysis of the productivity and citation impact of CLEF in the period 2010–2017 shows that CLEF is a very strong and vibrant initiative that has managed a major change of format between 2009/2010 and continues to produce relevant research, datasets and tools.

### 1 Introduction

The CLEF Initiative (Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum) has since 2000 been one of the major international initiatives that foster research, development and innovation of

---

B. Larsen (✉)

Science, Policy and Information Studies, Department of Communication and Psychology,  
Aalborg University, Copenhagen, Denmark  
e-mail: [birger@hum.aau.dk](mailto:birger@hum.aau.dk)

information access systems.<sup>1</sup> As CLEF approaches its 20th anniversary, we find it appropriate to study the scholarly impact of CLEF. We build on the analysis by Tsikrika et al. (2013) that studies the scholarly impact of CLEF Proceedings papers from the years 2000 to 2009. The present analysis focusses on the period 2010 to 2017, and the two studies together thus cover as much of CLEF as is possible up to the publication of this book. The end date of Tsikrika et al.'s study is appropriate as CLEF changed its format substantially in 2010. Prior to this, participants were invited to submit more detailed accounts of their approaches, expanded results and more in-depth analyses after each year's workshop and labs for the CLEF post-proceedings—which were peer-reviewed and published in Springer's Lecture Notes in Computer Science (LNCS) series. From 2010 onwards, the workshop and labs are accompanied by a 1-day peer-reviewed conference, and the proceedings of this coincide with that year's conference and labs. This has two major consequences: first, many fewer Working Notes and labs papers end up in the peer-reviewed proceedings, and secondly, the number of papers in the CLEF proceedings has dropped significantly. We choose therefore to include in the present study also the CLEF Working Notes that are published by each lab. We expect that the Working Notes papers will have markedly lower citation impact compared to the peer-reviewed Proceedings papers published in the highly visible and widely distributed Springer LNCS series.

The chapter is structured as follows: Following this introduction, we discuss our methodology, and then report on our findings. The chapter concludes with a discussion of the wider perspectives.

## 2 Methodology

As noted by Tsikrika et al. (2013) “*The scholarly impact of research activities is commonly measured by their associated publications (i.e., the publications generated as a result of such activities) and the citations they receive.*” (2013, p. 1). Studies of the impact of conferences are rare because conference proceedings are often poorly covered in citation indexes such as Clarivate Web of Science and Scopus. Google Scholar, which crawls the web for scientific papers often has much better coverage of publications in conference proceedings and the citations they receive. Tsikrika et al. (2013) compared the coverage of Scopus to Google Scholar and found that Google Scholar records almost ten times as many citations to the CLEF 2000–2009 proceedings compared to Scopus. We follow Tsikrika et al. (2013) and use Google Scholar in the present study. Google Scholar as a data source for bibliometric studies can be criticised—for an overview of this see Tsikrika et al. (2013). One of the shortcomings of Google Scholar is the lack of an ability to limit the citation windows, that is, to define the number of years after publication in which

---

<sup>1</sup><http://www.clef-initiative.eu/>.

citation data is collected. Without such a feature it is not possible to reproduce results of data collected at an earlier point in time—Google Scholar will always output the number of citations to a given paper from publication all the way up to the present. For this reason, we include also data on CLEF publications from 2009. This makes it possible to partially compare the results of the present study to those of Tsikrika et al. (2013), and to observe how large the increase of the number of citations is in Google Scholar after a number of years.

Google Scholar does not facilitate identification and direct download of data in batch mode for large sets of papers such as the 1000+ papers and notes published by CLEF between 2009 and 2017. A further challenge is that Google Scholar may not have comprehensive coverage of all CLEF papers. To create a comprehensive list of CLEF papers we used the DBLP Computer Science Bibliography.<sup>2</sup> We downloaded bibliographic data on all CLEF Proceedings and Working Notes to build a comprehensive list of CLEF papers for the period 2009–2017. We identified 218 papers from the CLEF Proceedings, and 1244 papers from the CLEF Working Notes in the period (Tables 1 and 2). Using header information from DBLP, we classified the CLEF Proceedings papers into the following paper types: *Frontmatter*, *Keynote*, *Overview*, *Panel*, *Full paper*, *Short paper* & *Best of Labs* (introduced in 2015; 1–2 extended papers from each lab selected by the lab organisers from the previous year’s Working Notes and subjected to peer-review).

As Google Scholar does not support batch mode download of large sets of papers we used Publish or Perish (PoP)—a software that, in response to a query, uses the Google Scholar API to retrieve up to 1000 publications and the number of citations received by each.<sup>3</sup> We used PoP to identify possible citations to the CLEF papers. With PoP we could issue queries to identify CLEF papers and download these in batches of up to 1000 papers. We issued a range of different queries, e.g. ‘CLEF’, ‘Cross-Language Evaluation Forum’, ‘Conference and Labs of the Evaluation Forum’, ‘CLEF Working Notes’ and variations thereof to cast a wide net to increase the chance that as many CLEF papers as possible were captured. A total of 2922 lines of data was retrieved using PoP, including duplicates across queries. The PoP data was then matched on year and title to the DBLP data using three approaches: (1) Exact match (approximately 66% of the DBLP data), (2) Partial match using IR best match techniques (17%), and (3) manual search directly in Google Scholar for the remaining papers (242 papers; 195 successfully identified). All data collection was carried out in March 2018 within a 7-day period. A total of 7 CLEF Proceedings papers (3%) and 47 CLEF Working Notes papers (4%) could not be identified in PoP or Google Scholar.

---

<sup>2</sup><http://dblp.uni-trier.de/>.

<sup>3</sup>Harzing, A.W. (2007) **Publish or Perish**, available from <https://harzing.com/resources/publish-or-perish>.

### 3 Results

The CLEF Proceedings contain a total of 218 papers during 2010–2017, between 16 and 51 each year. This is a marked drop from previous years (there were 133 Proceedings papers in 2009 and more than 100 papers per year 2005–2009 (Tsikrika et al. 2013, Table 1)). This is due to the change of format for CLEF where from 2010 onwards the Proceedings only publish a few select papers from a typically 1-day conference preceding the CLEF labs, the material of which is published in the Working Notes (Table 1). The CLEF Working Notes contain a total of 1244 papers 2010–2017, between 107 and 217 each year (Table 2). In 2009, before the change of format, this was 166 papers, and it seems likely that the number of Working Notes papers are comparable to the period before the format change.

The citation analysis shows that not every paper has received citations so far, but the proportion of cited papers is high—especially 3–4 years after a conference: 88–100% for the Proceedings and 80–90% for the Working Notes. It seems that the lag between publication and a high proportion of cited papers is longer for the Proceedings than Working Notes (e.g. 31% of the 2017 Proceedings papers are currently cited, compared to 60% of the Working Notes papers). The proportion of cited papers is also higher for the Working Notes papers (81% vs. 75% for the Proceedings). This is largely due to the slower citation rates for the Proceedings papers, with the latter generally reaching a higher proportion of cited papers after 3–4 years.

The mean number of citations per paper is very high, both for Proceedings papers (8.9) and Working Notes papers (7.5)—see also Fig. 1. It is interesting to observe that the gap between them is quite small—one might expect that the peer reviewed Proceedings papers published in the Springer Lecture Notes in Computer Science (LNCS) series would attract significantly more citations, but that is not the case. As expected the mean number of citations per paper drops the closer we get to the present with significantly fewer citations the last 3 years. It is also worth noting that the increase of citations to the early papers has grown dramatically: In April 2013, when the data for Tsikrika et al. (2013) was collected, the CLEF 2009 Proceedings

**Table 1** Publication and citation data for Proceedings papers 2009–2017

| Proceedings         | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | Total 2010–2017 |
|---------------------|------|------|------|------|------|------|------|------|------|-----------------|
| Papers              | 133  | 16   | 16   | 17   | 32   | 24   | 51   | 30   | 32   | 218             |
| Cited papers        | 121  | 14   | 15   | 17   | 31   | 23   | 37   | 17   | 10   | 164             |
| %Cited              | 91%  | 88%  | 94%  | 100% | 97%  | 96%  | 73%  | 57%  | 31%  | 75%             |
| Not found in GS     | 1    | 0    | 1    | 0    | 0    | 1    | 3    | 0    | 2    | 7               |
| Not found in GS (%) | 1%   | 0%   | 6%   | 0%   | 0%   | 4%   | 6%   | 0%   | 6%   | 3%              |
| Citations (GS)      | 1503 | 189  | 159  | 155  | 513  | 391  | 218  | 258  | 62   | 1945            |
| Citations per paper | 11.3 | 11.8 | 9.9  | 9.1  | 16.0 | 16.3 | 4.3  | 8.6  | 1.9  | 8.9             |
| Max. cites          | 131  | 30   | 24   | 42   | 139  | 94   | 38   | 91   | 17   |                 |

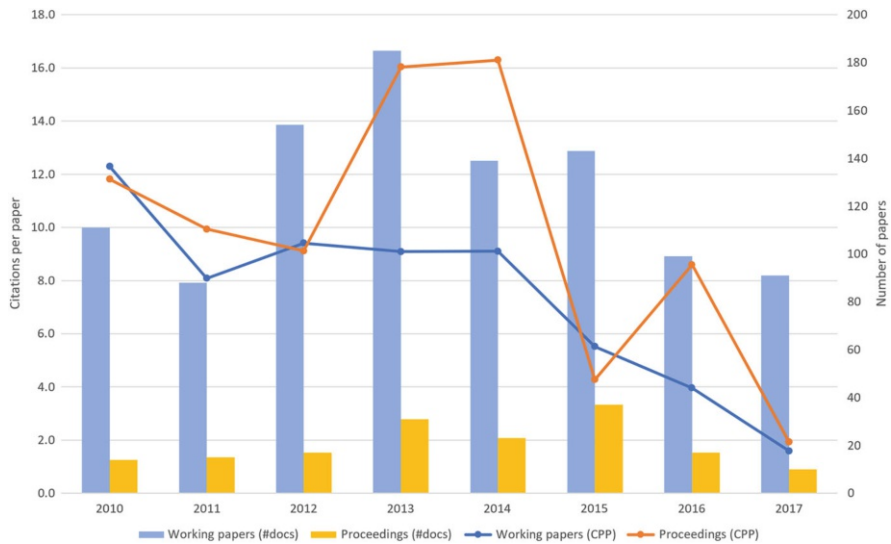
Source: DBLP, PoP and Google Scholar



**Table 2** Publication and citation data for Working Notes papers 2009–2017

| Working notes       | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | Total 2010–2017 |
|---------------------|------|------|------|------|------|------|------|------|------|-----------------|
| Papers              | 166  | 123  | 107  | 188  | 217  | 161  | 172  | 125  | 151  | 1244            |
| Cited papers        | 141  | 111  | 88   | 154  | 185  | 139  | 143  | 99   | 91   | 1010            |
| %Cited              | 85%  | 90%  | 82%  | 82%  | 85%  | 86%  | 83%  | 79%  | 60%  | 81%             |
| Not found in GS     | 6    | 4    | 8    | 8    | 7    | 6    | 3    | 2    | 9    | 47              |
| Not found in GS (%) | 4%   | 3%   | 7%   | 4%   | 3%   | 4%   | 2%   | 2%   | 6%   | 4%              |
| Citations (GS)      | 1569 | 1512 | 865  | 1769 | 1973 | 1466 | 949  | 495  | 241  | 9270            |
| Citations per paper | 9.5  | 12.3 | 8.1  | 9.4  | 9.1  | 9.1  | 5.5  | 4.0  | 1.6  | 7.5             |
| Max. cites          | 131  | 188  | 68   | 274  | 166  | 92   | 107  | 55   | 17   |                 |

Source: DBLP, PoP and Google Scholar



**Fig. 1** Number of Proceedings papers and Working note papers and their citation impact (mean citations per paper) for the period 2010–2017

papers had received a total of 770 citations in Google Scholar. In March 2018 the same 133 papers had almost doubled this to 1503 citations. The year 2015 stands out for the Proceedings papers with a marked drop in the mean number of citations per paper. This is mainly due to the inclusion of a large number of short papers in that year’s Proceedings (20 out of 51), which tend to have lower citation rates and thus lower the average. In addition, we may note that the full papers of 2015 also have received quite few citations so far (see Tables 5 and 6).

Tables 1 and 2 also show the number of citations of the most cited Proceedings paper and Working Notes paper respectively in a given year. As expected there are large differences in these, with the most highly cited Proceedings paper receiving 139 citations and the most cited Working Notes paper receiving 274 citations.

**Table 3** Top 10 cited Proceedings papers (2010–2017)

| Year | Title   | Citations |
|------|---|-----------|
| 2013 | Overview of the ShARe/CLEF eHealth Evaluation Lab 2013                          | 139       |
| 2014 | Improving the Reproducibility of PAN’s Shared Tasks: Plagiarism Detection       | 94        |
| 2013 | Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems        | 93        |
| 2016 | LifeCLEF 2016: Multimedia Life Species Identification Challenges                | 91        |
| 2014 | Overview of the ShARe/CLEF eHealth Evaluation Lab 2014                          | 73        |
| 2016 | General Overview of ImageCLEF at the CLEF 2016 Labs                             | 65        |
| 2013 | Multilingual Question Answering over Linked Data (QALD-3): Lab Overview         | 52        |
| 2014 | Overview of RepLab 2014: Author Profiling and Reputation Dimensions for...      | 51        |
| 2013 | Recent Trends in Digital Text Forensics and Its Evaluation—Plagiarism Detection | 43        |
| 2012 | Bringing the Algorithms to the Data: Cloud-Based Benchmarking for Medical...    | 42        |

**Table 4** Top 10 cited Working Note papers (2010–2017)

| Year | Title  | Citations |
|------|--|-----------|
| 2012 | Overview of the 4th International Competition on Plagiarism Detection        | 274       |
| 2010 | Overview of the 1st International Competition on Wikipedia Vandalism...      | 188       |
| 2013 | Overview of the 5th International Competition on Plagiarism Detection        | 166       |
| 2010 | Overview of the 2nd International Competition on Plagiarism Detection        | 143       |
| 2013 | Overview of the Author Profiling Task at PAN 2013                            | 117       |
| 2013 | Overview of the ImageCLEF 2013 Medical Tasks                                 | 108       |
| 2015 | Overview of the 3rd Author Profiling Task at PAN 2015                        | 107       |
| 2012 | Overview of the ImageCLEF 2012 Medical Image Retrieval and Classification... | 100       |
| 2014 | ISOFT at QALD-4: Semantic Similarity-based Question Answering System...      | 92        |
| 2013 | The ImageCLEF 2013 Plant Identification Task                                 | 86        |

Tables 3 and 4 list the top 10 most cited Proceedings and Working Notes papers respectively. Notably, almost all of them are overview papers. This makes sense as they are the logical papers to cite when using CLEF data for research—both in the Working Notes papers of that year’s conference and subsequently when publishing work based on CLEF data.

For the Proceedings papers we can meaningfully divide them into publication types. Table 5 shows the number of publications across types, and Table 6 the mean number of citations per paper for these. We see some changes in which types of publications are included over the years: All years have varying number of *Full Papers* (7–22), *Overview Papers* are introduced into the Proceedings 2013 onwards, *Best of Labs Papers* from the previous year from 2015 onwards, and *Short Papers* are mainly included 2015 onwards. Table 6 shows that *Overview Papers* generally have a high citation impact—in particular the *Overview Papers* of 2013 and 2014 have attracted a high number of citations (Fig. 1), and several of them are in the top 10 most cited (Table 3). The *Full Papers* show a quite high impact, which declines steadily as we get closer to the present, as expected. The remaining paper types have a relatively low citation impact on average.

**Table 5** Number of publications across Proceedings paper types 2009–2010

|              | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | Total 2010–2017 |
|--------------|------|------|------|------|------|------|------|------|------|-----------------|
| Best of labs |      |      |      |      |      |      | 8    | 5    | 6    | 19              |
| Frontmatter  | 1    |      |      |      |      |      |      |      |      |                 |
| Full-paper   | 124  | 12   | 14   | 14   | 22   | 16   | 15   | 10   | 7    | 110             |
| Keynote      |      | 2    | 2    |      |      |      |      |      |      | 4               |
| Overview     | 8    |      |      |      | 10   | 8    | 8    | 7    | 10   | 43              |
| Panel        |      | 2    |      |      |      |      |      |      |      | 2               |
| Short        |      |      |      | 3    |      |      | 20   | 8    | 9    | 40              |
| Total        | 133  | 16   | 16   | 17   | 32   | 24   | 51   | 30   | 32   | 218             |

**Table 6** Mean number of citations per paper across Proceedings paper types 2009–2010

|              | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | Total 2010–2017 |
|--------------|------|------|------|------|------|------|------|------|------|-----------------|
| Best of labs |      |      |      |      |      |      | 5.3  | 0.6  | 0.3  | 2.5             |
| Frontmatter  | 30.0 |      |      |      |      |      |      |      |      |                 |
| Full-paper   | 8.2  | 15.2 | 11.3 | 10.6 | 5.7  | 5.3  | 2.4  | 2.9  | 0.1  | 6.9             |
| Keynote      |      | 1.5  | 0.5  |      |      |      |      |      |      | 1.0             |
| Overview     | 57.1 |      |      |      | 38.8 | 38.4 | 11.8 | 32.0 | 5.9  | 24.9            |
| Panel        |      | 2.0  |      |      |      |      |      |      |      | 2.0             |
| Short        |      |      |      | 2.3  |      |      | 2.3  | 0.3  | 0.0  | 1.4             |
| Total        |      | 11.8 | 9.9  | 9.1  | 16.0 | 16.3 | 4.3  | 8.6  | 1.9  | 8.9             |

## 4 Discussion and Conclusion

The analysis of the productivity and citation impact of CLEF in the period 2010–2017 shows that CLEF is a very strong and vibrant initiative that has managed a major change of format between 2009/2010 and that continues to produce relevant research, datasets and tools. This bibliometric analysis is the first to include the CLEF lab Working Notes which show interesting results together with those of the more formal Proceedings papers. Significantly lower citation impact might be expected from the Working Notes: they are non-peer reviewed, have a work-in-progress-nature, are more numerous (typically more than five times as many per year than Proceedings papers). However, the analysis of Google Scholar citations shows that the Working Notes papers on average obtain a citation impact that is almost on par with the Proceedings papers (7.5 versus 8.9; only 17.6% lower). It is worth noting that although a slightly larger proportion of Proceedings papers are cited at least once, the Working Notes papers seem to be cited earlier—that is, sooner after publication. A possible explanation could be that the Working Notes papers are open access and freely available on the Web. Overall, the working papers are a major contribution to the impact of CLEF in absolute numbers—with 9270 citations to Working Notes papers and 1945 citations to Proceedings papers in the period.

The lab overview papers are very dominant among the cited papers (Tables 3 and 4): among the Proceedings papers (where we can determine the publication types) they account for less than one fifth of the publications, but receive more than half of the citations. From the titles of the most cited Working Notes papers (Table 4) it seems likely that this is also the case for the Working Note papers. One might speculate that the high citation rates of the overview papers is due to citations from other CLEF papers from the same year. We cannot check this easily on a paper by paper basis with our current data, but such CLEF citations cannot account for the high citation rates—there are simply not enough papers every year to explain that many citations.

The CLEF Proceedings papers during 2000–2009 are covered in Tsikrika et al. (2013). The present analysis covers the CLEF Proceedings as well as Working Notes papers during 2009–2017. The scholarly impact of two sets of publications remain to be analysed: The impact of the Working Notes during 2000–2008, and the impact of those publications called ‘CLEF-derived publications’ by Tsikrika et al. (2013)—e.g. the impact of journal articles based on CLEF data. An analysis of the latter could yield interesting insight into the impact of the datasets that have been generated in the CLEF labs and made available to the research community. We leave the analysis of these two sets of publications for future research.

**Acknowledgements** We wish to thank Lucas Chaves, University of Copenhagen, Denmark for assistance in matching DBLP and Google Scholar data, Anne-Wil Harzing for help with the ‘Publish or Perish’ software platform, as well as two anonymous reviewers for constructive feedback.

## References

- Tsikrika T, Larsen B, Müller H, Endrullis S, Rahm E (2013) The scholarly impact of CLEF (2000–2009). In: Forner P, Müller H, Paredes R, Rosso P, Stein B (eds) Information access evaluation meets multilinguality, multimodality, and visualization. Proceedings of the fourth international conference of the CLEF initiative (CLEF 2013). Lecture notes in computer science (LNCS), vol 8138. Springer, Heidelberg, pp 1–12

# Reproducibility and Validity in CLEF



Norbert Fuhr

**Abstract** In this paper, we investigate CLEF’s contribution to the reproducibility of IR experiments. After discussing the concepts of reproducibility and validity, we show that CLEF has not only produced test collections that can be re-used by other researchers, but also undertaken various efforts in enabling reproducibility.

## 1 Introduction

Reproducibility of experiments is an important concept in research, supporting validation of reported results as well as allowing for later comparison with new approaches. The ACM task force on reproducibility stated: “A scientific result is not fully established until it has been independently reproduced”.<sup>1</sup>

Evaluation initiatives like e.g. CLEF, TREC, FIRE, NTCIR support reproducibility by sharing experimental resources—in contrast to research based on proprietary data that cannot be reproduced by other researchers (due to e.g. copyright or privacy issues).

In this paper, we take a closer look at the concept of reproducibility and to what extent it has been supported in different CLEF tracks over the years. Section 2 sketches a core model of reproducibility that was developed recently, followed by a section that briefly discusses internal validity of IR experiments. In Sect. 4 we investigate reproducibility in different types of CLEF tracks. Finally, Sect. 5 summarizes the findings and points out some issues for further research.

---

<sup>1</sup><https://www.acm.org/data-software-reproducibility>.

N. Fuhr (✉)  
University of Duisburg-Essen, Duisburg, Germany  
e-mail: [norbert.fuhr@uni-due.de](mailto:norbert.fuhr@uni-due.de)

## 2 Models of Reproducibility

The PRIMAD (pronounce “primed”) model of reproducibility was developed during a Dagstuhl Seminar (Freire et al. 2016; Ferro et al. 2016). It describes a framework for specifying the major components of an experiment:

- Research Goal characterizes the purpose of a study;
- Method refers to the specific approach proposed or considered by the researcher;
- Implementation relates to the actual implementation of the method (usually in some programming language);
- Platform describes the underlying hard- and software like the operating system and the computer used;
- Data comprises both the input data as well as the specific parameters chosen to carry out the method;
- Actor is the experimenter.

The term PRIMAD is derived from the first letters of the component names, but in a different order: **P**latform—**R**esearch goal—**I**mplementation—**M**ethod—**A**ctor—**D**ata. In the following, we use these letters to characterize specific forms of reproducibility.

As an example, consider a student performing a retrieval experiment. The research goal is to achieve a high retrieval quality, and the method chosen is the BM25 formula. Experiments use the TERRIER system as implementation, under the operating system Ubuntu 16.04 on a Dell xyz server. The GOV2 collection serves as input data, and a specific setting of the BM25 parameters is chosen. The actor is the student performing the runs.

When another researcher tries to reproduce this experiment, she will change one or more of the components. If she tries to rerun the experiment without changing anything else, then we have another actor, that is, A is changed to A', the actor is “primed”. If successful, this experiment would demonstrate that the original researcher has supplied enough information to ensure reproducibility. If the results of the experiment are the same, then the original findings have been successfully reproduced and thus confirmed.

Now let us look at changes of the other components, which are more interesting:

$R \rightarrow R'$ : When the research goal is changed, then we *repurpose* some of the components of the experiment for another research question (for example, performing interactive retrieval experiments). So method and implementation usually are also changed.

$M \rightarrow M'$ : Most of the research in the field of IR deals with the investigation of alternative methods (retrieval models, formulas). This implies also a new implementation  $I'$ , possibly running on a different platform. However, for performing comparisons, the (input) data should be the same.

$I \rightarrow I'$ : Here a researcher uses a different implementation, say Lemur instead of Terrier, or does her own reimplementaion.

- $P \rightarrow P'$ : In most cases, independent researchers do not have access to the platform used in the original experiment. Even different versions of system libraries might have subtle effects on the outcome of experiments.
- $D \rightarrow D'$ : Rerunning an experiment with different parameters might be useful for testing the robustness of a method. Applying the implementation to different input data (for example, test collections) aims at investigating the generality of the method.
- $A \rightarrow A'$ : While changing the actor might not be very interesting in system-oriented approaches, this may become relevant if we are dealing with user experiments where the experimenter interacts with the test subjects.

In order to ensure reproducibility, there is the need to be able to share as many PRIMAD components as possible. Research goal and method are what we currently share via publications in conference proceedings or journals (although details of the method are often missing); for deep learning methods, however, a text-only description of the actual model used can never be sufficient, it should be handled like an implementation: Sharing an implementation is possible via making it open source and uploading it on Web sites focusing on this task (for example, Github). Platforms can be shared by means of virtual machines or dockers, or by “evaluation as a service”. For the input data, there are a number of standard test collections which are generally available. When researchers use their own test collection, however, reproducibility can only be ensured when this collection is shared with the community, ideally via a trustworthy repository.

Finally, there are two other important aspects that are not part of the core PRIMAD model:

**Transparency** is the ability to look into all necessary components to verify that the experiment does what it claims; for example, sharing a virtual machine, but not the source code of an experiment, would not satisfy this criterion.

**Consistency** refers to the success or failure of a reproducibility experiment in terms of consistent outcomes; for example, using a random number generator for breaking ties in a ranking would lead to problems with respect to this criterion—thus experiments should be designed in a way that avoids these problems. In science, there usually is some knowledge about the precision of the measurement devices employed; thus, one can tell whether or not two different numbers represent consistent outcomes. In IR, we lack this kind of knowledge.

### 3 Validity

Since IR experiments are stochastic experiments, the concept of consistency may be too strict, and also miss an important point. A more suitable concept is **internal**

**validity**, which is described in Wikipedia<sup>2</sup> as “...*the extent to which a causal conclusion based on a study is warranted, which is determined by the degree to which a study minimizes systematic error (or ‘bias’)*. It contrasts with external validity, the degree to which it is warranted to generalize results to other contexts.”

For example, if an experiment uses the simple holdout method for separation between training and testing data, then the results are highly dependent on the actual split, and different splits will show a high variance of outcomes (see e.g. Rao et al. 2015). Just reproducing the original results by using the same split might demonstrate consistency, but fail to address (internal) validity. A better approach would be to use  $k$ -fold cross validation (see e.g. Witten et al. 2011, pp. 152–6) already in the original experiment (or doing this even  $k$  times with different partitionings). In this case, reproducing the experiment might not yield exactly the same figures (and thus fail to show consistency), but the numbers should be rather similar, and thus the causal conclusion should be the same. If confidence intervals for the measured values were computed (and published), a reproducibility experiment with other data from the same population yielding results within this intervals would also validate the original findings.

Fuhr (2017) points out some other evaluation practices that hamper internal validity. Here we want to mention just the two most important areas:

- Choice of evaluation metric: Metrics like MRR or ERR are theoretically invalid, and MAP is based on assumptions that are often inappropriate for the task studied.
- Multiple testing: When more than one significance test is performed on the same data set, then the significance levels have to be corrected subject to the number of tests (e.g. Bonferroni’s method divides the desired  $p$ -value by the number of hypotheses in order to get the significance level to test on) (Carterette 2012). Alternatively, one can apply a post-hoc test such as for example Tukey’s, which implicitly considers all pairwise comparisons (e.g. between all the runs submitted for a track) (Braschler 2002).

Of course, the ultimate goal of IR research is to achieve external validity. However, there is little research addressing this issue. Frequently, authors apply a method to different test collections, in order to demonstrate that it achieves internal validity on all of them. Thus, the implicit claim is that the new method is universally valid—which is unscientific, and hardly ever true. It would be more interesting to have a clear statement about the applicability of a method: What are the underlying assumptions, and how can we check them on a new data set (without having to perform actual retrieval tests)?

---

<sup>2</sup>[https://en.wikipedia.org/wiki/Internal\\_validity](https://en.wikipedia.org/wiki/Internal_validity).



## 4 Reproducibility in CLEF

### 4.1 *Standard Test Collections*

Most CLEF tracks follow the scheme that was originally developed in TREC: Each track focuses on a specific research goal, for which a test collection is provided. The participating groups apply their own methods, using their own implementation and platform and running the experiments themselves. So, in terms of the PRIMAD model, everything but the **R**esearch goal and the **D**ata is primed (strictly speaking, the parameters and the output data differ from run to run). Of greatest interest for the participating researchers are the rankings of runs. In order to avoid the multiple testing problem mentioned above, Tukey's test has been used regularly in the early years of CLEF (Braschler 2002). However, the knowledge about this method seems to have been forgotten, and so we have seen several cases of multiple testing without correction in recent years.

A persisting problem weakening internal validity is the application of the holdout method, which is immanent to the design of tracks, by releasing the testing data only after the training of the methods is finished. Moreover, determining the 'ground truth' for this testing data often depends on the output of the participating runs (usually via some pooling method). Thus, switching the roles of testing and training data, or applying cross-validation might not be possible. However, for tracks where we have comparable data for training and testing samples, an investigation into the variance of results would be very helpful.

In many tracks, we have two dimensions of data, namely the 'documents' and the 'tasks' (e.g. retrieval topics). While the discussion from above mainly relates to the documents dimension, varying the task set can often be performed more easily. Voorhees and Buckley (2002) presents a study of this kind for the TREC ad-hoc track.

Test collections are a valuable resource for many research groups (even those not participating in the original track), who mostly follow the same research goal. However, as pointed out in Carterette (2012), the more a test collection is used, the higher is the likelihood that a new method might outperform previous methods just by chance; this is both due to the (usually ignored) problems of multiple testing as well as of sequential testing (using knowledge about the outcome of previous tests before formulating the hypothesis).

As we focus on reproducibility here, it is more interesting to ask if and how runs have been reproduced by other research groups. Obviously, since all research groups are working on the same test collections, and outputs of the participating runs are generally available, there seems to be little need to actually reproduce the original runs. A notable exception is the study presented in Armstrong et al. (2009), where a large number of methods on the TREC ad-hoc collection was rerun. In this work (and others of the same type), the goal was to use the original **M**ethod, but reimplement it and run it on a different platform. Armstrong et al. (2009) found that most methods used poor baselines for comparison, and were in fact not able to

improve on the best known results for these collections; moreover, when combining the various methods, the improvements (observed over the poor baselines) did not add up.

## 4.2 *Evaluation as a Service*

In some of the PAN tracks (Potthast et al. 2012), a different approach to reproducibility was taken: Instead of sharing only the data, participants also had access to the evaluation platform; thus they had to deliver an implementation of their method that was runnable on the common platform. In some tracks (Potthast et al. 2013), the evaluation infrastructure also provided a search API that was accessible for the implementation, thus making sure that certain details of the method/implementation were common for all participants. The latter issue is important for many IR experiments, as basic functions like tokenization, stemming or the stopword list are hardly ever specified in publications, but might effect results.

So evaluation as a service supports reproducibility by providing the same Platform and easing the sharing of the Implementation. However, only when the source program is disclosed, transparency is also ensured.

A special variant of evaluation as a service are living labs (Schuth et al. 2015). Here participants provide an Implementation that is run in an operational system (again on a common Platform), where (implicit) feedback from real users are collected. The obvious advantage is that the performance of the submitted method is evaluated with real life tasks in an online environment. On the other hand, reproducing these results later becomes more difficult, since the data usually is not static, and the functionality of the system (starting with subtle changes of the user interface) is subject to continuous improvement; technically, it is possible to archive the database state and the system version. However, user needs and their expectations might change over time.

## 4.3 *Interactive Retrieval*

Starting with the interactive cross-lingual question answering track in 2004 (Gonzalo and Oard 2005), various tracks investigated interactive retrieval. The usual setup was A/B tests in the form of laboratory experiments, either with the same two systems used by all participants, or by a baseline system and a participant-specific system. The two systems implemented different methods, where the differences could be either in the user interface only, or also in the underlying retrieval methods. Usually, both the underlying test collection as well as the log data of all test subjects was published at the end of the track.

The living labs approach described above is a specific instance of interactive retrieval (although the research focus was on the system side).

There has been little research on reproducibility of interactive IR (see also the description of challenges in reproducing this type of experiments in Ferro et al. 2016). In system-oriented approaches, one uses an implementation (either the one from the original experiment or a new one, possibly for a new method), run it on a test collection and then compare the output with the ground truth. For interactive retrieval, however, one needs new test subjects. Even if the data and the implementation of the original system were available, there is still the problem of recruiting test subjects who are comparable to those of the original study. Furthermore, other context factors might have to be considered.

As an approach for avoiding some of these problems, the INFILE track (Besançon et al. 2010) used automatic interactive feedback for simulating information filtering. This method works under the assumption that we have complete relevance information for a set of queries (which has been collected before via system pooling). Reproducibility in this case is comparable to that of system-oriented experiments; however, user interaction is limited to relevance feedback.

In a recent study in psychology (Open Science Collaboration 2015), less than half of the results from 100 experiments published in top journals could be replicated. Besides improper evaluation methods, the underspecified test designs also posed problems in reproducing the original experiments. Most fundamental, however, is the publication bias: Usually only positive test results get accepted for publication, while studies with negative results get rejected; thus, a large fraction of the published work is based on random results. If e.g. 20 researchers investigate non-existing effects, one of them will observe a difference at the 95% significance level, and will publish this result, while the other 19 will go on and look for other effects.

## 5 Towards Better Reproducibility

The discussion above shows that CLEF has made substantial contributions to enable reproducibility of experiments. A cornerstone in this effort is the evaluation infrastructure in the form of the DIRECT system (Di Nunzio and Ferro 2005; Agosti et al. 2012; Silvello et al. 2017) that manages the data produced in the various evaluation campaigns of CLEF. A major limitation of this system, however, is the fact that it only deals with the output data. For the test collections themselves, there is no uniform solution (partly due to the various restrictions attached to most of these collections), but they are usually accessible for later investigations, also for other researchers. However, keeping track of the use of these testbeds would also be very fruitful, in order to deal with the multiple or sequential testing problem; this way, one could view experimental results on re-used collections in the proper context.

Looking at the PRIMAD model, one can see that there have been limited efforts in sharing implementations and platforms—mostly only during the campaign itself, while later studies might be difficult to impossible in most cases. Thus, it would

be helpful to set up a repository for this purpose, using current containerization technology (like e.g. Docker) for encapsulating complete experiments, so that they can easily be rerun by other researchers.

The reproducibility of living labs has already been recognized as a problem within CLEF (Kille et al. 2017), but needs more research.

Internal validity of experiments is also an important issue. As pointed out in Fuhr (2017), one can find serious experimental errors in published papers of major IR venues, and even some evaluation campaigns use flawed procedures. Here track organizers of evaluation campaigns should put more emphasis on using a proper evaluation methodology.

With respect to the publication bias mentioned above, the CLEF Working notes are already a nice exception, as they contain both positive and negative results. However, researchers should be encouraged to look deeper into negative results, especially when a seemingly good idea does not lead to the expected outcome. Of course, program committees and journal editorial boards must become more open towards accepting also papers with negative results.

## 6 Conclusion and Outlook

Reproducibility and internal validity are important properties of experimental studies. For a long time, only minor attention was paid to these issues, but the situation has changed in recent years. Evaluation initiatives like CLEF play an important role in raising the scientific quality of experiments: While the major goal of a track usually is the investigation of certain types of IR tasks, a track also defines the experimental methodology to be used for evaluation. As shown in this paper, the various CLEF tracks have introduced experimental standards and also ensured reproducibility in many ways, although there are still several open issues. Thus future organizers should consider reproducibility and internal validity as essential criteria when defining a track.

On a more general level, our field should aim at establishing standards and infrastructures supporting reproducibility. This is also a problem for conferences and journals in our field which should put more emphasis on the reproducibility of the results published.

Finally, external validity of experimental results still remains an open issue: How can we generalize the findings of a set of experiments towards other data sets and application domains? With the rich set of experimental results collected from the various CLEF campaigns, it is possible to perform metastudies (Angelini et al. 2016) and formulate some general observations. More systematic research in this direction is needed, however (Ferro et al. 2018).

## References

- Agosti M, Di Buccio E, Ferro N, Masiero I, Peruzzo S, Silvello G (2012) DIRECTIONS: design and specification of an IR evaluation infrastructure. In: Catarci T, Forner P, Hiemstra D, Peñas A, Santucci G (eds) Information access evaluation. Multilinguality, multimodality, and visual analytics. Proceedings of the third international conference of the CLEF initiative (CLEF 2012). Lecture notes in computer science (LNCS), vol 7488. Springer, Heidelberg, pp 88–99
- Agostini M, Ferro N, Santucci G, Silvello G (2016) A visual analytics approach for what-if analysis of information retrieval systems. In Perego R, Sebastiani F, Aslam JA, Ruthven I, Zobel J (eds) Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval, SIGIR 2016, Pisa, July 17–21, 2016. ACM, New York, pp 1081–1084. ISBN 978-1-4503-4069-4. <http://doi.acm.org/10.1145/2911451.2911462>
- Armstrong TG, Moffat A, Webber W, Zobel J (2009) Improvements that don't add up: ad-hoc retrieval results since 1998. In: Cheung DW-L, Song I-Y, Chu WW, Hu X, Lin JJ (eds) Proceedings of the 18th ACM conference on Information and knowledge management CIKM. ACM, New York, pp 601–610. ISBN 978-1-60558-512-3
- Besaçon R, Chaudiron S, Mostefa D, Timimi I, Choukri K, Laïb M (2010) Information filtering evaluation: overview of CLEF 2009 INFILE track. In: Peters C, Di Nunzio GM, Kurimo M, Mandl T, Mostefa D, Peñas A, Roda G (eds) Multilingual information access evaluation vol. I. Text retrieval experiments – tenth workshop of the cross–language evaluation forum (CLEF 2009). Revised selected papers. Lecture notes in computer science (LNCS), vol 6241. Springer, Heidelberg, pp 342–353
- Braschler M (2002) CLEF 2001 – overview of results. In: Peters C, Braschler M, Gonzalo J, Kluck M (eds) Evaluation of cross-language information retrieval systems: second workshop of the cross–language evaluation forum (CLEF 2001) revised papers. Lecture notes in computer science (LNCS), vol 2406. Springer, Heidelberg, pp 9–26
- Carterette BA (2012) Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Trans Inf Syst* 30(1):4:1–4:34. <http://doi.acm.org/10.1145/2094072.2094076>
- Di Nunzio GM, Ferro N (2005) DIRECT: a system for evaluating information access components of digital libraries. In: Rauber A, Christodoulakis C, Tjoa AM (eds) Research and advanced technology for digital libraries, 9th European conference, ECDL 2005, Vienna, Austria, September 18–23, 2005, proceedings. Springer, Berlin, pp 483–484. [https://doi.org/10.1007/11551362\\_46](https://doi.org/10.1007/11551362_46)
- Ferro N, Fuhr N, Jarvelin K, Kando N, Lippold M, Zobel J (2016) Increasing reproducibility in IR: findings from the Dagstuhl seminar on “reproducibility of data-oriented experiments in e-science”. *SIGIR Forum* 50(1):68–82. <http://sigir.org/files/forum/2016J/p068.pdf>
- Ferro N, Fuhr N, Grefenstette G, Konstan JA, Castells P, Daly EM, Declerck T, Ekstrand MD, Geyer W, Gonzalo J, Kuflik T, Linden K, Magnini B, Nie J-Y, Perego R, Shapira B, Soboroff I, Tintarev N, Verspoor K, Willemsen MC, Zobel J (2018) The Dagstuhl perspectives workshop on performance modeling and prediction. *SIGIR Forum* 52(1):91–101
- Freire J, Fuhr N, Rauber A (2016) Reproducibility of data-oriented experiments in e-science. *Dagstuhl Rep* 6(1):108–159. [http://drops.dagstuhl.de/opus/institut\\_dagrep.php?fakultaet=07](http://drops.dagstuhl.de/opus/institut_dagrep.php?fakultaet=07)
- Fuhr N (2017) Some common mistakes in ir evaluation, and how they can be avoided. *SIGIR Forum* 51(3):32–41. <http://sigir.org/wp-content/uploads/2018/01/p032.pdf>
- Gonzalo J, Oard DW (2005) iCLEF 2004 track overview: pilot experiments in interactive cross-language question answering. In: Peters C, Clough P, Gonzalo J, Jones GJF, Kluck M, Magnini B (eds) Multilingual information access for text, speech and images: fifth workshop of the cross–language evaluation forum (CLEF 2004) revised selected papers. Lecture notes in computer science (LNCS), vol 3491. Springer, Heidelberg, pp 310–322

- Kille B, Lommatzsch A, Hopfgartner F, Larson M, Brodt T (2017) CLEF 2017 newsreel overview: offline and online evaluation of stream-based news recommender systems. In Cappellato L, Ferro N, Goeuriot L, Mandl T (eds) Working notes of CLEF 2017 - conference and labs of the evaluation forum, Dublin, September 11–14, 2017. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1866/>. [http://ceur-ws.org/Vol-1866/invited\\_paper\\_17.pdf](http://ceur-ws.org/Vol-1866/invited_paper_17.pdf)
- Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science* 349(6251):943–952
- Potthast M, Gollub T, Hagen M, Kiesel J, Michel M, Oberländer A, Tippmann M, Barrón-Cedeño A, Gupta P, Rosso P, Stein B (2012) Overview of the 4th international competition on plagiarism detection. In: Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) CLEF 2012 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073, <http://ceur-ws.org/Vol-1178/>
- Potthast M, Hagen M, Gollub T, Tippmann M, Kiesel J, Rosso P, Stamatatos E, Stein B (2013) Overview of the 5th international competition on plagiarism detection. In: Forner P, Navigli R, Tufis D, Ferro N (eds) CLEF 2013 working notes. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1179/>
- Rao J, Lin JJ, Efron M (2015) Reproducible experiments on lexical and temporal feedback for tweet search. In Hanbury A, Kazai G, Rauber A, Fuhr N (eds) Advances in information retrieval - 37th European conference on IR research, ECIR 2015, Vienna, March 29–April 2, 2015. Proceedings. Lecture Notes in Computer Science, vol 9022, pp 755–767. ISBN 978-3-319-16353-6. [https://doi.org/10.1007/978-3-319-16354-3\\_82](https://doi.org/10.1007/978-3-319-16354-3_82)
- Schuth A, Balog K, Kelly L (2015) Overview of the living labs for information retrieval evaluation (LL4IR) CLEF Lab 2015. In: Mothe J, Savoy J, Kamps J, Pinel-Sauvagnat K, Jones GJF, SanJuan E, Cappellato L, Ferro N (eds) Experimental IR meets multilinguality, multimodality, and interaction. Proceedings of the sixth international conference of the CLEF association (CLEF 2015). Lecture notes in computer science (LNCS), vol 9283. Springer, Heidelberg, pp 484–496
- Silvello G, Bordea G, Ferro N, Buitelaar P, Bogers T (2017) Semantic representation and enrichment of information retrieval experimental data. *Int J Digit Libr* 18(2):145–172. ISSN 1432-5012. <https://doi.org/10.1007/s00799-016-0172-8>
- Voorhees EM, Buckley C (2002) The effect of topic set size on retrieval experiment error. In: Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '02. ACM, New York, pp 316–323. ISBN 1-58113-561-0. <https://doi.org/10.1145/564376.564432>
- Witten IH, Frank E, Hall MA (2011) Data mining: practical machine learning tools and techniques, 3rd edn. Morgan Kaufmann, San Francisco. ISBN 0123748569, ISBN 9780123748560

# Visual Analytics and IR Experimental Evaluation



Nicola Ferro and Giuseppe Santucci

**Abstract** We investigate the application of *Visual Analytics (VA)* techniques to the exploration and interpretation of *Information Retrieval (IR)* experimental data. We first briefly introduce the main concepts about VA and then we present some relevant examples of VA prototypes developed for better investigating IR evaluation data. Finally, we conclude with a discussion of the current trends and future challenges on this topic.

## 1 Visual Analytics

Around the year 2000, in order to support human beings in analyzing large and complex datasets, synergies between *Information Visualization (IV)* and *Data Mining (DM)* started to be considered. *Visual Data Mining (VDM)* was defined as a new area focused on the explorative analysis of visually represented data. In 2001, the first VDM workshop was held in Freiburg. In 2004, first in the United States, and almost at the same time in Europe, researchers started talking about Visual Analytics (Wong and Thomas 2004). Unlike VDM, there is the clear intention to focus on the analysis process that leads to explanation, interpretation, and presentation of hidden information in the data, taking advantage of dynamic visualizations. From that moment on, the term VDM was superseded by the term *Visual Analytics (VA)*. Daniel Keim, one of the major European experts in the field, provides the following definition: “Visual analytics is more than just visualization

---

N. Ferro (✉)

Department of Information Engineering, University of Padua, Padova, Italy

e-mail: [ferro@dei.unipd.it](mailto:ferro@dei.unipd.it)

G. Santucci

Department of Computer, Control, and Management Engineering “Antonio Ruberti”, Sapienza

University of Rome, Rome, Italy

e-mail: [santucci@dis.uniroma1.it](mailto:santucci@dis.uniroma1.it)

© Springer Nature Switzerland AG 2019

N. Ferro, C. Peters (eds.), *Information Retrieval Evaluation*

*in a Changing World*, The Information Retrieval Series 41,

[https://doi.org/10.1007/978-3-030-22948-1\\_24](https://doi.org/10.1007/978-3-030-22948-1_24)

and can rather be seen as an integrated approach combining visualization, human factors and data analysis”.

On a grand scale, VA provides technology that combines the strengths of human and electronic data processing. Visualization becomes the medium of a semi-automated analytical process, where humans and machines cooperate using their respective distinct capabilities for the most effective results. The user has to be the ultimate authority in giving the direction of the analysis along his or her specific task. At the same time, the system has to provide effective means of interaction to concentrate on this specific task since in many applications different people work along the path from data to decision.

Figure 1 schematizes the VA process that combines automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data. The figure shows an abstract overview of the different stages (represented through ovals) and their transitions (arrows) in the VA process.

The first step is often to preprocess and transform the data to derive different representations for further exploration (as indicated by the Transformation arrow). Other typical preprocessing tasks include data cleaning, normalization, grouping, or integration of heterogeneous data sources. After the transformation, the analyst may choose between applying visual or automatic analysis methods. Alternating between visual and automatic methods is characteristic for the VA process and leads to a continuous refinement and verification of preliminary results. User interaction with the visualization is needed to reveal insightful information, for instance by zooming in on different data areas or by considering different visual views on the data. In summary, in the VA process, knowledge can be gained from visualization and automatic analysis, as well as the preceding interactions between visualizations, models, and the human analysts. With respect to the field of visualization, VA integrates methodology from Information Visualization (Card et al. 1999; Chen 2004; Spence 2007; Ware 2012), Visual Data Mining (Keim 2001), geospatial

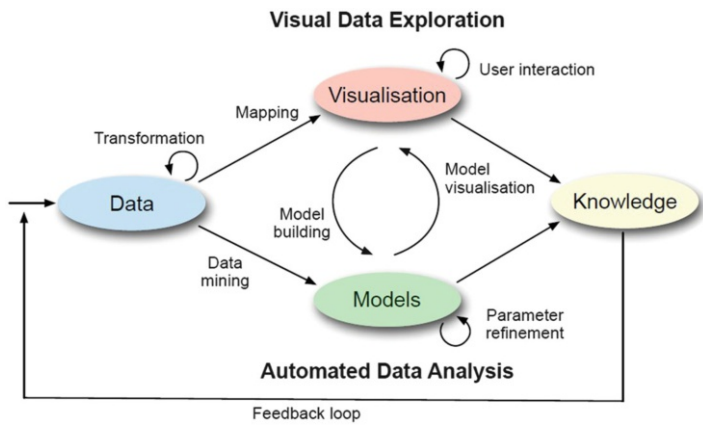
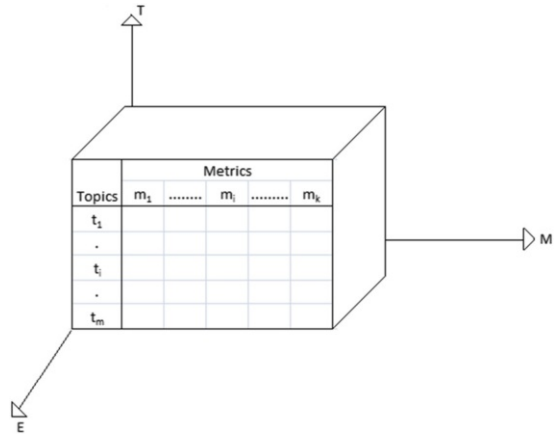


Fig. 1 The visual analytics process (Keim et al. 2010)



**Fig. 2** The overall TME data cube with the  $TM(e)$  transformation highlighted



analytics (Andrienko et al. 2007), and scientific analytics. In particular, human factors (e.g., interaction, cognition, perception, collaboration, presentation, and dissemination) play a key role in the communication between human and computer, as well as in the decision-making process, see, e.g., Keim et al. (2006).

## 2 The IR Evaluation Data Cube

As shown in Fig. 1, the initial step of any analysis is to get a clear understanding of the data involved in the process, in our case the data used within IR evaluation. Despite the strong differences that exist among the different domains targeted by IR applications, IR systems are typically evaluated according to the common Cranfield paradigm (Cleverdon 1967), which allows us to compare the effectiveness of different IR systems on the same collection. The scientific data produced during evaluation are then arranged across several transformations that are suitable for different analysis patterns. In the European Union project PROMISE<sup>1</sup> these data plus their transformations have been formalized as follows.

The initial view on the data is represented by the *Topics–Metrics–Experiments (TME)* data cube, shown in Fig. 2, reporting for each experiment (i.e., an IR system) its performance according to different evaluation measures across a set of topics.

Starting from this cube, it is possible to transform data in different ways, according to different analysis objectives. In particular four kinds of transformations have been identified.

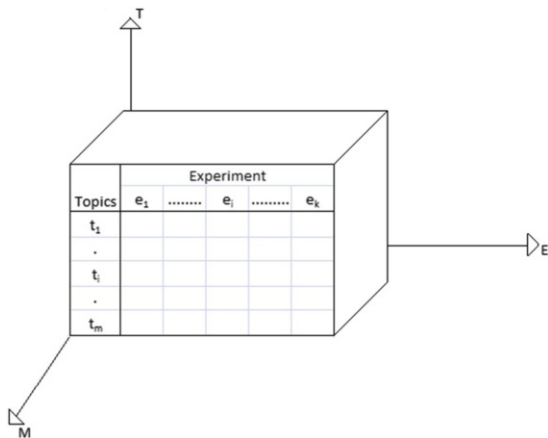
The first kind of transformation makes it possible to analyze the performance of a single experiment  $e$ , i.e. an IR system, with respect to topics and it is the projection

<sup>1</sup><http://www.promise-noe.eu/>.

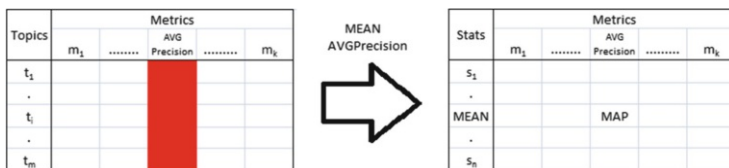
of the TME cube on the Topics–Metrics axes of experiment  $e$ . In particular, this table is a matrix  $T \times M$ , where  $T$  is the set of topics and  $M$  is the set of metrics. In the following, we refer to this kind of transformation as TM( $e$ ) tables (Topics  $\times$  Metrics table of experiment  $e$ , shown in Fig. 2).

A second kind of transformation, shown in Fig. 3, is useful to analyze the behavior of a set of experiments, i.e. IR systems, over a set of topics with respect to a single metric  $m$ , which is the most common case in IR evaluation. In particular, this table is represented by a  $T \times E$  matrix where  $T$  is the set of topics and  $E$  is the set of experiments. In the following, we refer to this kind of transformation TE( $m$ ) tables (Topics  $\times$  Experiments table of metric  $m$ , shown in Fig. 2). Comparisons are made along rows, to evaluate the behavior of a single topic, or among columns to compare two or more experiments.

The third kind of transformation describes a single experiment  $e$  in terms of descriptive statistics computed over a set of topics with respect to different metrics. In particular, this table is represented by an  $S \times M$  matrix where  $S$  is the set of descriptive statistics and  $M$  is the set of metrics. In the following, we refer to this kind of transformation as the SM( $e$ ) table (Statistics  $\times$  Metrics table of experiment  $e$ , shown in Fig. 4). This table is strictly related to the corresponding TM( $e$ ) table since values are computed from the TM( $e$ ) table columns. Figure 4



**Fig. 3** Projection of the TME data cube on the topics-experiments axes with the TE( $m$ ) transformation



**Fig. 4** Relationship between TM and SM tables

shows an example of how a  $TM(e)$  table can be used to calculate values of the  $SM(e)$  table.

As shown in Fig. 4, in an  $SM(e)$  table there is the same number of metrics as in the corresponding  $TM(e)$  table. If we extend this table with respect to experiments, we obtain a new cube, the *Statistics–Metrics–Experiments* ( $SME$ ) data cube, shown in Fig. 5. With respect to the  $SME$  cube, an  $SM(e)$  table is a projection on the Statistics–Metrics axes.

The last kind of table we consider, allows us to inspect a single metric  $m$  in terms of descriptive statistics and experiments, i.e., it makes it possible to compare different experiments against some descriptive statistics computed on a given metric. In particular, this table is represented by an  $S \times E$  matrix where  $S$  is the set of statistics and  $E$  is the set of experiments. In the following, we refer to this transformation as the  $SE(m)$  table (Statistics  $\times$  Experiment table computed on metric  $m$ , shown in Fig. 6) and it is a projection of the  $SME$  cube on the Statistics–Experiments axes.

Fig. 5 The  $SME$  data cube

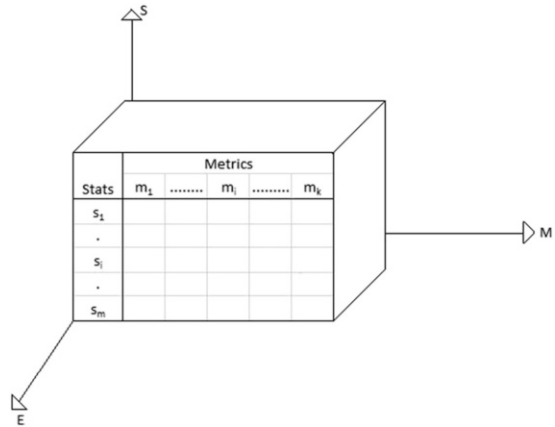
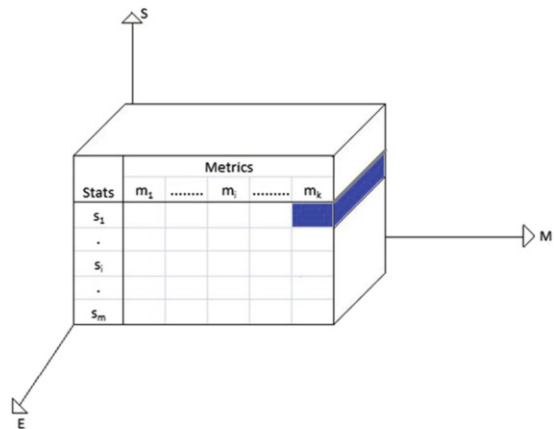


Fig. 6 The  $SME$  data cube projected on the statistics-experiments axes



As discussed above, all these data and their transformations constitute the entry step depicted in the leftmost part of Fig. 1.

### 3 Examples of VA Systems on the IR Evaluation Data Cube

In this section, we present some recent examples of systems which exploit VA techniques to improve IR experimental evaluation and to analyse and interact with IR experimental data. They represent different types of instantiations of the “Models” and “Visualisation” steps depicted in Fig. 1.

#### 3.1 VAIRĚ

Angelini et al. (2017) presented a VA environment, called *Visual Analytics for Information Retrieval Evaluation (VAIRĚ)*, which uses multiple visualizations working on different aspects of the data. Visualizations are synchronized using two main interaction mechanisms: *selection* (a way to focus the attention on a subset of data) and *highlight* (it allows to highlight a part of the displayed data maintaining the context). IR evaluation data cube transformations are then mapped to multiple coordinated visualizations.

Moreover, considering that user activities are quite repetitive and follow several basic analysis patterns, VAIRĚ provides some ad-hoc, highly automated patterns for analysis: *Per topic analysis* and *Per Experiment analysis*.

The system supports six visualizations, listed from the simplest to the most advanced: bi-dimensional scatter-plots, stacked bar-charts, box plots, table lens, enhanced frequency distribution, and the Precision-Recall-chart, all of them particularly suited for evaluation tasks in IR. Depending on the chosen type of analysis, the system will present the user with different subsets of these visualizations. Nonetheless, the user can customize the environment by simply removing a visualization and dragging a new one from a menu.

*Per topic analysis* it makes it possible to compare a set of experiments on each topic with respect to a chosen evaluation measure. Therefore the first step for a user is to select an evaluation measure  $m$ . Looking at the TME data cube described in the previous section, we can note that choosing an evaluation measure is equivalent to fixing an axis and reducing the set of data to the  $TE(m)$  transformation. Per topic analysis implies a comparison on each topic, so, by default, we represent topics on the x-axis in each available visualization. We provide four views for a per topic analysis: table lens, a boxplot chart, a scatter plot, and a stacked bar chart.

The user can change the evaluation measure under analysis and restrict her/his focus on data subsets through select and highlight operations. As an example, Fig. 7 shows three topics highlighted in all the four visualizations.



Fig. 7 Per topic analysis: an highlight operation

*Per experiment analysis* it makes it possible to analyze an experiment as a whole and/or compare the performance of a set of experiments with respect to a chosen descriptive statistics. As an example, on Fig. 8, left side, the table represents an experiment in each row, showing the descriptive statistics of *Average Precision (AP)* (min, max, median, etc.). The box plot chart (McGill et al. 1978) in Fig. 8, right side, shows the percentile values of the observed metric for each experiment represented through boxplots.

### 3.2 VIRTUE

Figure 9 shows the overall framework of *Visual Information Retrieval Tool for Upfront Evaluation (VIRTUE)* to support the evaluation workflow (Angelini et al. 2014): *performance analysis* and *failure analysis* are the traditional phases carried

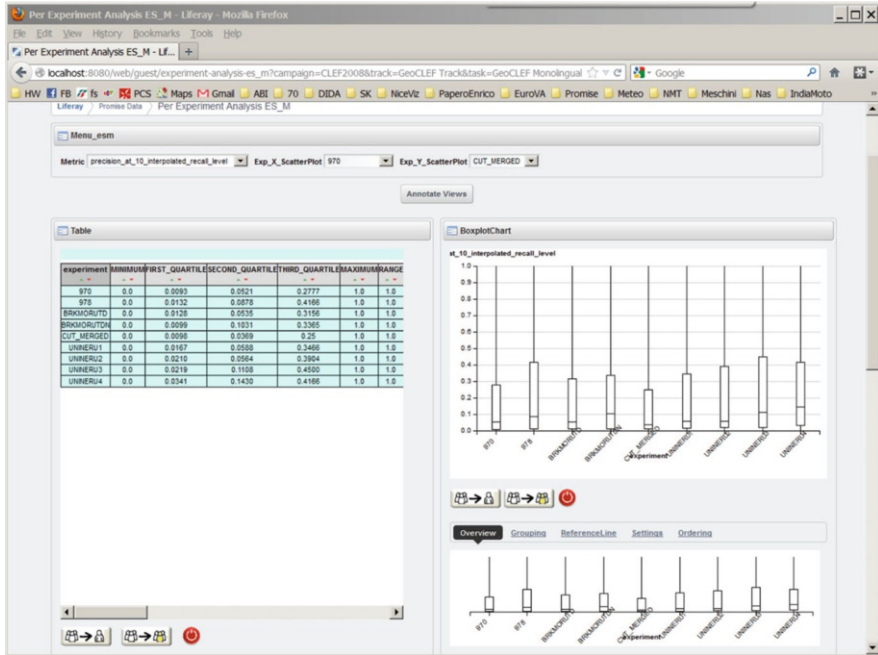


Fig. 8 Per experiment analysis: table and box plot

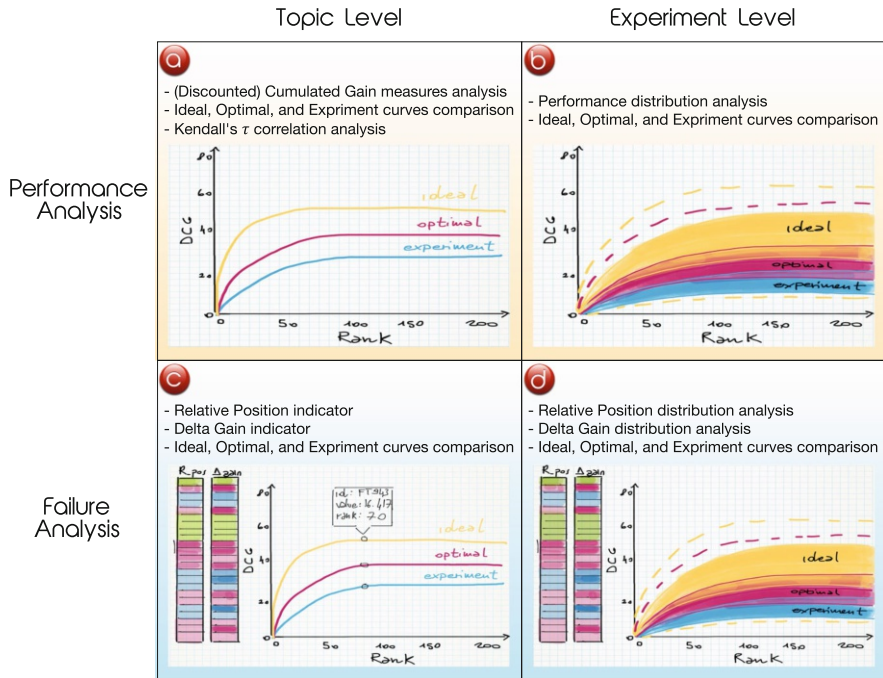
out during experimental evaluation, where VIRTUE contributes to make them more effective and to reduce the needed effort via both tailored visualizations and high interaction with the experimental data.

*Topic Level* concerns the analysis of the documents retrieved in response to a given topic of a run while *Experiment Level* deals with overall statistics and effects concerning the whole set of topics of a run, i.e., all the different ranked lists of retrieved documents.

In both the topic and experiment level analyses, the user is presented with three curves, reporting the *Discounted Cumulated Gain (DCG)* (Järvelin and Kekäläinen 2002) in three cases: (a) the actual performance (experiment curve), (b) the improvement that is possible to achieve reordering the actual result in the optimal way (optimal curve), and (c) the best possible score, in which the results contain *all* the relevant documents in the optimal way (ideal curve). On the leftmost part, two bars represent the ranked list of retrieved documents where colors in the leftmost bar indicate how much a document has been misplaced with respect to its ideal position in the ranking and colors in the rightmost bar indicate the gain loss in terms of DCG due to this misplacement.

Therefore, VIRTUE:

- supports performance analysis on a topic-by-topic basis and with aggregate statistics over the whole set of topics;



**Fig. 9** VIRTUE overall framework. (a) Ranked results exploration. (b) Ranked results distribution exploration. (c) Failing documents identification. (d) Failing topics identification

- facilitates failure analysis to allow researchers and developers to more easily spot and understand failing documents and topics.

The main target users of VIRTUE are domain experts, i.e., researchers and developers in the IR and related fields who need to understand and improve their systems. Moreover, VIRTUE can also be useful for educational purposes, e.g. in undergraduate or PhD courses where information retrieval is taught and where explaining how to interpret the performances of an IR system is an important part of the teaching. Finally, it may also find application in production contexts as a tool for monitoring and interpreting the performances of a running system so as to ensure that the desired service levels are met.

### 3.3 VATE<sup>2</sup>

*Visual Analytics Tool for Experimental Evaluation (VATE<sup>2</sup>)* Angelini et al. (2012, 2016b,a) introduced a new phase in the evaluation workflow, called *what-if analysis*. It falls between the experimental evaluation and the design and implementation of the identified modifications. What-if analysis aims at estimating what the effects of

a modification to the IR system under examination could be, before actually being implemented. In this way researchers and developers can get a feeling of whether a modification is worth being implemented and, if so, they can go ahead with its implementation followed by a new evaluation and analysis cycle for understanding whether it has produced the expected outcomes.

What-if analysis exploits VA techniques to make researchers and developers: (1) interact with and explore the ranked result list produced by an IR system and the achieved performances; (2) hypothesize possible causes of failure and their fixes; (3) estimate the possible impact of such fixes through a powerful analytical model of the system behavior.

Figure 10 shows the mock-up used for designing the VATE<sup>2</sup> user interface whose objective is to provide a rough estimation of what could be the impact of fixing a possible failure on the performances in order to assess if it might be worth implementing it or not. What visualization of Fig. 10 offers to the user is: (1) the possibility of dragging and dropping the target document in the desired position of the rank; (2) the estimation of which other documents would be affected by the movement of the target document and how the overall ranking would be modified; (3) the computation of the system performances according to the new ranking. Therefore, moving a single target document would actually cause the movement and repositioning of a whole set of documents that share features impacted by the same modification which will affect the target document selected by the user. These complex interactions between documents may generate modifications on

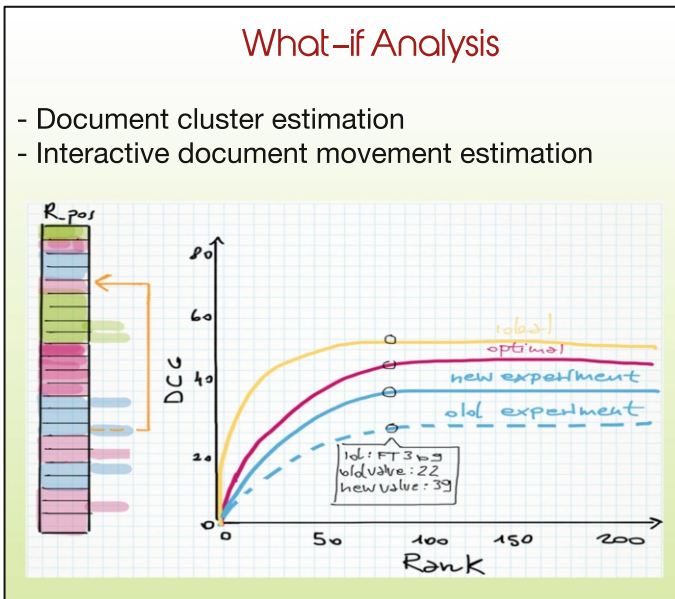


Fig. 10 VATE<sup>2</sup> overview



the ranking that go well beyond what the user imagined when moving the single target document and which are definitely hard for her/him to guess. Thus, the contribution of the visualization and analytical engine of Fig. 10 is to automatically point out to the user all these complex interactions and how they affect the overall ranking.

Once the new ranked list has been produced by using a clustering and movement strategy, the performances of this new ranked list are computed and the corresponding new line is shown to the user so that he can assess whether the hypothesized modification may be beneficial or not. In the former case VATE<sup>2</sup> turns on a green light to indicate to the user that s/he should go on with the fix of the system, otherwise it turns on a red light meaning that the fix may be useless or worsen the system.

### 3.4 The RETRIEVAL Online Platform

Ioannakis et al. (2018) developed RETRIEVAL,<sup>2</sup> a Web-based integrated platform for performance evaluation of IR methods, which shares many commonalities with the VAIRĚ system discussed in Sect. 3.1.

RETRIEVAL allows users to upload their datasets in various formats, converting them into internal data structures which resemble the IR evaluation data cube we described in Sect. 2. RETRIEVAL supports different evaluation measures, like AP (Buckley and Voorhees 2005), *normalized Discounted Cumulative Gain (nDCG)* (Järvelin and Kekäläinen 2002), *Rank-Biased Precision (RBP)* (Moffat and Zobel 2008), and many others.

Once the data cube has been created, RETRIEVAL provides several alternative visualisations, shown in Fig. 11, such as a precision-recall graph (Fig. 11c), a scatter-plot where each pixel indicates a relevant/not relevant document (Fig. 11g), a dissimilarity matrix map where the user can identify a normalized dis-similarity distance between any two items using an interactive pointer that offers real-time zoom-in functionality (Fig. 11e), a tabular view of the data (Fig. 11d), and more.

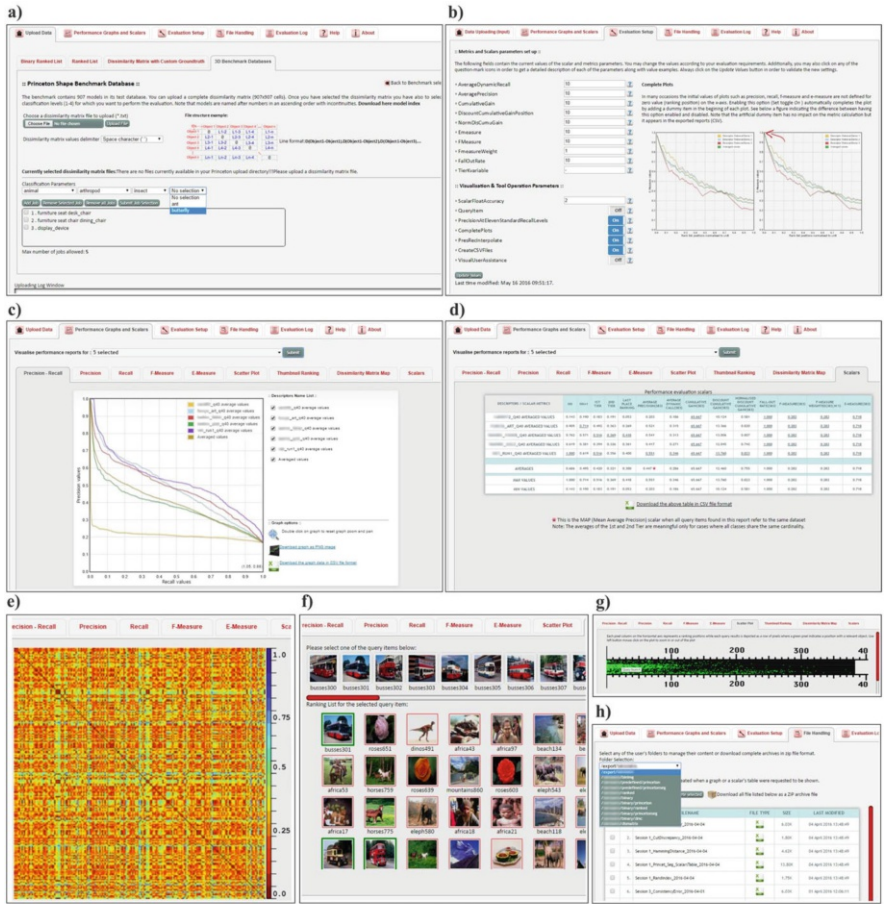
### 3.5 The Visual Pool System

Lipani et al. (2017) proposed Visual Pool<sup>3</sup> an IV system to explore alternative pooling strategies to build the ground truth of a test collection.

---

<sup>2</sup><http://retrieval.ceti.gr/>.

<sup>3</sup><http://visualpool.aldolipani.com/>.



**Fig. 11** Example of the RETRIEVAL user interface (Ioannakis et al. 2018). Downloaded from the RETRIEVAL Facebook page (<https://www.facebook.com/RetrievalEvaluationTool/>). (a) PSB batch evaluation interface. (b) Performance metrics parameterisation. (c) Precision–recall curves. (d) Scalar metrics table. (e) Dissimilarity matrix visualisation. (f) Thumbnail-based ranking. (g) Binary relevance scatter plot. (h) User’s file repository

Figure 12 shows the user interface of Visual Pool. Users can load a set of runs, which are displayed in the left part of the window where each column is a system and each row is a retrieved document. The topmost left button allows users to select among different pooling strategies, whose effects are then interactively displayed. Moreover, users can load an already existing set of relevance judgments whose statistics are reported in the middle of the window. The color coding is as follows with respect to the loaded relevance judgments: red is for not relevant documents; green is for relevant documents; gray is for not pooled documents; and, black is for pooled documents which are not contained in the currently loaded relevance

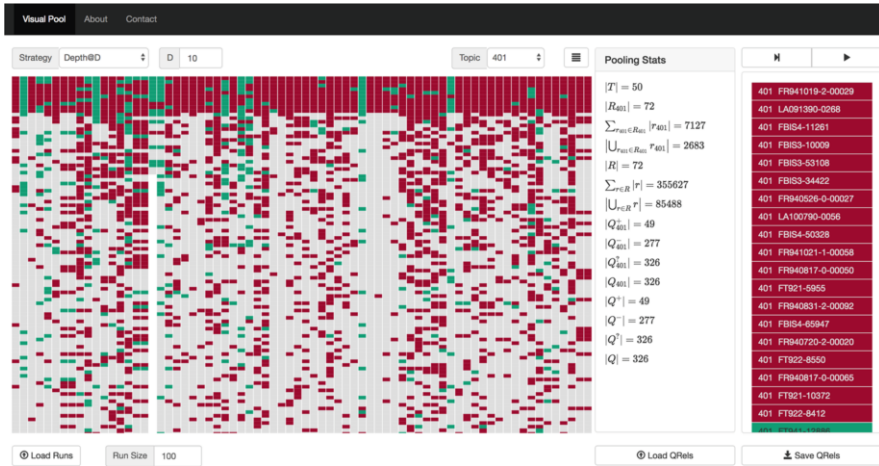


Fig. 12 Example of the Visual Pool user interface (Lipani et al. 2017). Courtesy of Aldo Lipani

judgments. Finally, the rightmost part of the window shows the details of the currently loaded systems and of the pooling method.

Overall, Visual Pool allows users to interactively experiment alternative pooling strategies over a set of runs, compare their effects with respect to an existing set of relevance assessments, and to assess their intrinsic bias.

### 3.6 CLAIRE

Angelini et al. (2018) developed *Combinatorial visual Analytics system for Information Retrieval Evaluation (CLAIRE)*,<sup>4</sup> a VA system for exploring and making sense of the performances of a large number of IR systems, in order to quickly and intuitively grasp which system configurations are preferred, what are the contributions of the different components and how these components interact together. In particular, CLAIRE allows users to explore, analyze, interact with a *Grid of Points (GoP)* (Ferro and Harman 2010), i.e. a very large set of IR systems originated from all the possible combinations of targeted components—stop lists, stemmers, and IR models in the case of Fig. 13.

The goal of CLAIRE is to avoid the need for complex statistical analyses, such as those based on *ANalysis Of VAriance (ANOVA)* by Ferro and Silvello (2016), while fostering a more natural and intuitive way of making sense of such set of systems.

<sup>4</sup><http://awareserver.dis.uniroma1.it:11768/claire/>.

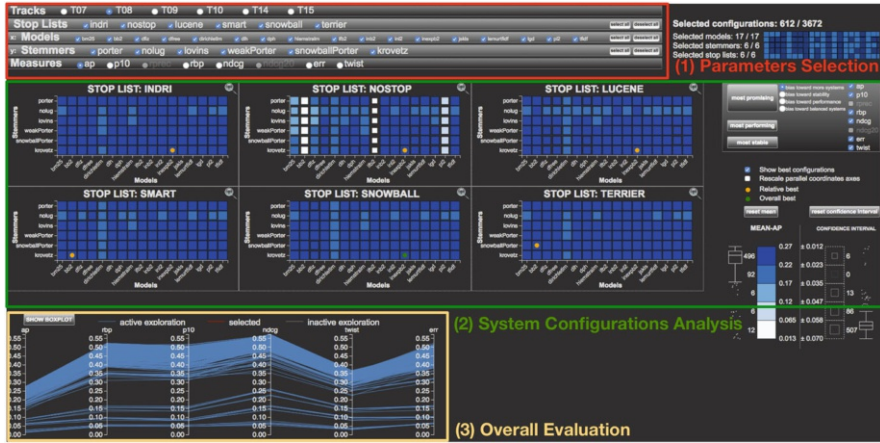


Fig. 13 Example of the CLAIRE user interface (Angelini et al. 2018)

Figure 13 shows the user interface of CLAIRE:

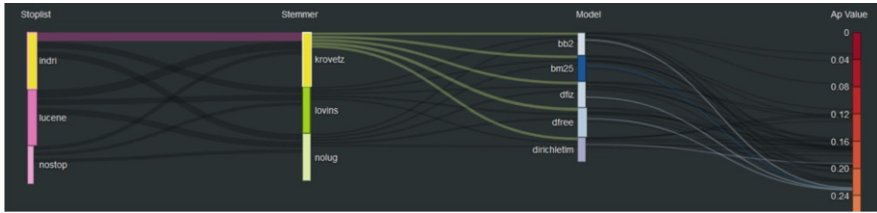
1. The *Parameters Selection* area deals with the exploration coordinates, i.e., collections, stop lists, stemmers, IR models, and evaluation measures;
2. The *System Configurations Analysis* area enables the performance analysis of the system configurations using a specific evaluation measure. The multidimensional performance space is mapped to a bidimensional one by using a set of tiles where the color and the size of the tiles represent, respectively, the average performance and the confidence interval for that performance;
3. The *Overall Evaluation* area, where the system configurations performances are evaluated across the complete set of evaluation measures by using a parallel coordinates plot (Inselberg 2009).

CLAIRE relies on the multiple coordinated views design, which allows users to propagate the results of the analysis process steps among all these three areas.

### 3.7 Sankey GoP

Rocco and Silvello (2019) further investigated how to intuitively explore and make sense of a GoP by leveraging a Sankey diagram (Sankey 1898; Schmidt 2008).

As shown in Fig. 14, Rocco and Silvello replaced the tile-based visualization of CLAIRE with a Sankey diagram which makes it possible to represent the multidimensional performance space as a flow of performance from one component to another in the pipeline constituting an IR system. A single system is represented by a path, i.e. a series of links connecting one component with the next one. The user can select a set of components to highlight the paths of interest. The component



**Fig. 14** Example of the Sankey GoP user interface (Rocco and Silvello 2019). Courtesy of Gianmaria Silvello

columns present a number of rectangles equal to the components selected in the parameter selection area and the size of the rectangle gives a visual idea of the performances of the component it represents.

## 4 Discussion and Challenges

IV and VA techniques have been traditionally exploited mostly for the presentation and exploration of the results returned by an IR system (Zhang 2008). The purpose of these components is to increase the ability to fulfill IR tasks where visualization is the natural platform for browsing and query searching. Some examples are: identification of the objects and their attributes to be displayed (Fowler et al. 1991); different ways of presenting the data (Morse et al. 2002); the definition of visual spaces and visual semantic frameworks (Zhang 2001); using rankings for presenting the user with the most relevant visualizations (Seo and Shneiderman 2005), for browsing the ranked results (Derthick et al. 2003), or for comparing large sets of rankings (Behrisch et al. 2013). The development of interactive means for IR is an active field which focuses on search user interfaces (Hearst 2009, 2011), displaying of results (Crestani et al. 2004) and browsing capabilities (Koshman 2005).

In the context of IR evaluation, IV strategies have been adopted for analyzing experimental runs, e.g. beadplots in (Banks et al. 1999). Each row in a beadplot corresponds to a system and each “bead”, which can be gray or colored, corresponds to a document. The position of the bead across the row indicates the rank position in the result list returned by the system. The same color indicates the same document and therefore the plot makes it easy to identify a group of documents that tend to be ranked near to each other and to compare the performance of different systems. As a further example, *Query Performance Analyzer (QPA)* (Sormunen et al. 2002) provides the user with an intuitive idea of the distribution of relevant documents in the top ranked positions through a relevance bar, where rank positions of the relevant documents are highlighted, and it also allows for the comparison between the Recall-Precision graphs of a query and the most effective query formulations issued by users for the same topic.

Nevertheless, much less attention has been generally devoted to applying VA techniques to the analysis and exploration of the performance of IR systems in order to get a better understanding of their behaviour, when and where they fail, and how to improve them.

In Sect. 3 we have presented some recent examples which start to explore how VA can be applied to improve the IR evaluation workflow and to better interact, analyse, interpret, and understand the performance of IR systems.

We can consider the examples discussed in Sect. 3 as positive indicators of a rising interest for this topic in the research community, even if the full potential of VA for IR evaluation is still far from being fully unfledged.

Moreover, designing and developing this kind of systems is still extremely challenging because they require not only very specialist competence in both fields—IR and VA—but also a good mutual understanding of what are the main issues, approaches, and techniques in both fields. This sort of cross-disciplinary competencies and reciprocal interest in exploring each other's field is not easy to find. Moreover, joint collaborations must be established between research groups operating in the two fields and willing to invest in something which may be perceived as not mainstream in both fields.

Overall, we think that IR can greatly benefit from using and developing VA techniques to enhance and ease the exploration of the experimental results in order to build better systems. Moreover, the visual interpretation and understanding of IR system performance might even be considered as a community goal in the same way as the explicability and interpretability of IR algorithms is now perceived as a more and more compelling need. On the other hand, IR can be a very relevant domain for VA researchers, especially considering its pervasiveness in daily life. Indeed, IR evaluation poses challenges in terms of the complexity and the huge amount of the data to be analysed as well as the sophistication of the statistical methods used to make sense of the data. Finally, the increasing use of traces for capturing and predicting user behavior is adding a new complexity layer to the whole process, making the call for VA in IR louder.

## References

- Andrienko G, Andrienko N, Jankowski P, Keim DA, Kraak MJ, MacEachren A, Wrobel S (2007) Geovisual analytics for spatial decision support: setting the research agenda. *Int J Geogr Inf Sci* 21(8):839–858
- Angelini M, Ferro N, Santucci G, Silvello G (2012) Visual interactive failure analysis: supporting users in information retrieval evaluation. In: Kamps J, Kraaij W, Fuhr N (eds) *Proceedings of 4th symposium on information interaction in context (IIIX 2012)*. ACM Press, New York, pp 195–203
- Angelini M, Ferro N, Santucci G, Silvello G (2014) VIRTUE: a visual tool for information retrieval performance evaluation and failure analysis. *J Vis Lang Comput* 25(4):394–413
- Angelini M, Ferro N, Santucci G, Silvello G (2016a) A visual analytics approach for what-if analysis of information retrieval systems. In: Perego R, Sebastiani F, Aslam J, Ruthven I,

- Zobel J (eds) Proceedings of 39th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2016). ACM Press, New York, pp 1081–1084
- Angelini M, Ferro N, Santucci G, Silvello G (2016b) What-if analysis: a visual analytics approach to information retrieval evaluation. In: Di Nunzio GM, Nardini FM, Orlando S (eds) Proceedings of 7th Italian information retrieval workshop (IIR 2016). CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073. <http://ceur-ws.org/Vol-1653/>
- Angelini M, Ferro N, Santucci G, Silvello G (2017) Visual analytics for information retrieval evaluation campaigns. In: Sedlmair M, Tominski C (eds) Proceedings of 8th international workshop on visual analytics (EuroVA 2017). Eurographics Association, Goslar, pp 25–29
- Angelini M, Fazzini V, Ferro N, Santucci G, Silvello G (2018) CLAIRE: a combinatorial visual analytics system for information retrieval evaluation. *Inf Process Manag* 54(6), 1077–1100
- Banks D, Over P, Zhang NF (1999) Blind men and elephants: six approaches to TREC data. *Inf Retrieval* 1(1–2):7–34
- Behrisch M, Davey J, Simon S, Schreck T, Keim D, Kohlhammer J (2013) Visual comparison of orderings and rankings. In: Pohl M, Schumann H (eds) Proceedings of 4th international workshop on visual analytics (EuroVA 2013). Eurographics Association, Goslar
- Buckley C, Voorhees EM (2005) Retrieval system evaluation. In: Harman DK, Voorhees EM (eds) TREC. experiment and evaluation in information retrieval. MIT Press, Cambridge, pp 53–78
- Card SK, Mackinlay JD, Shneiderman B (1999) Readings in information visualization: using vision to think. Morgan Kaufmann Publishers, San Francisco, CA
- Chen C (2004) Information visualization - beyond the horizon. Springer, London
- Cleverdon CW (1967) The cranfield tests on index languages devices. *Aslib Proc* 19(6):173–194
- Crestani F, Vegas J, de la Fuente P (2004) A graphical user interface for the retrieval of hierarchically structured documents. *Inf Process Manag* 40(2):269–289
- Derthick M, Christel MG, Hauptmann AG, Wactlar HD (2003) Constant density displays using diversity sampling. In: Munzner T, North S (eds) Proceedings of 9th IEEE symposium on information visualization (INFOVIS 2003). IEEE Computer Society, Los Alamitos, pp 137–144
- Ferro N, Harman D (2010) CLEF 2009: Grid@CLEF pilot track overview. In: Peters C, Di Nunzio GM, Kurimo M, Mandl T, Mostefa D, Peñas A, Roda G (eds) Multilingual information access evaluation vol. I text retrieval experiments – tenth workshop of the cross-language evaluation forum (CLEF 2009). Revised selected papers. Lecture notes in computer science (LNCS), vol 6241. Springer, Heidelberg, pp 552–565
- Ferro N, Silvello G (2016) A general linear mixed models approach to study system component effects. In: Perego R, Sebastiani F, Aslam J, Ruthven I, Zobel J (eds) Proceedings of 39th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2016). ACM Press, New York, pp 25–34
- Fowler RH, Lawrence-Fowler WA, Wilson BA (1991) Integrating query, thesaurus, and documents through a common visual representation. In: Fox EA (ed) Proceedings of 14th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 1991). ACM Press, New York, pp 142–151
- Hearst MA (2009) Search user interfaces, 1st edn. Cambridge University Press, New York
- Hearst MA (2011) “Natural” search user interfaces. *Commun ACM* 54(11):60–67
- Inselberg A (2009) Parallel coordinates. Visual multidimensional geometry and its applications. Springer, New York
- Ioannakis G, Koutsoudis A, Pratikakis I, Chamzas C (2018) Retrieval—an online performance evaluation tool for information retrieval methods. *IEEE Trans Multimedia* 20(1):119–127
- Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of IR techniques. *ACM Trans Inf Syst* 20(4):422–446
- Keim DA (2001) Visual exploration of large data sets. *Commun ACM* 44(8):38–44
- Keim DA, Mansmann F, Schneidewind J, Ziegler H (2006) Challenges in visual data analysis. In: Banissi E (ed) Proceedings of the 10th international conference on information visualization (IV 2006). IEEE Computer Society, Los Alamitos, pp 9–16

- Keim DA, Kohlhammer J, Ellis G, Mansmann F (eds) (2010) *Mastering the information age – solving problems with visual analytics*. Eurographics Association, Goslar
- Koshman S (2005) Testing user interaction with a prototype visualization-based information retrieval system. *J Am Soc Inf Sci Technol* 56(8):824–833
- Lipani A, Lupu M, Hanbury A (2017) Visual pool: a tool to visualize and interact with the pooling method. In: Kando N, Sakai T, Joho H, Li H, de Vries AP, White RW (eds) *Proceedings of 40th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2017)*. ACM Press, New York, pp 1321–1324
- McGill R, Tukey JW, Larsen WA (1978) Variations of box plots. *Am Stat* 32(1):12–16
- Moffat A, Zobel J (2008) Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans Inf Syst (TOIS)* 27(1):2:1–2:27
- Morse EL, Lewis M, Olsen KA (2002) Testing visual information retrieval methodologies case study: comparative analysis of textual, icon, graphical, and spring displays. *J Am Soc Inf Sci Technol* 53(1):28–40
- Rocco G, Silvello G (2019) An InfoVis tool for interactive component-based evaluation. *arXiv.org, information retrieval (csIR)* arXiv:1901.11372
- Sankey HR (1898) Introductory note on the thermal efficiency of steam-engines. Report of the committee appointed on the 31st March, 1896, to consider and report to the council upon the subject of the definition of a standard or standards of thermal efficiency for steam-engines: with an introductory note. *Minutes of proceedings of the institution of civil engineers*, vol 134, pp 278–283 including Plate 5
- Schmidt M (2008) The sankey diagram in energy and material flow management. *J Ind Ecol* 12(1):82–94
- Seo J, Shneiderman B (2005) A rank-by-feature framework for interactive exploration of multidimensional data. *Inf Vis* 4(2):96–113
- Sormunen E, Hokkanen S, Kangaslampi P, Pyy P, Sepponen B (2002) Query performance analyser – a web-based tool for IR research and instruction. In: Järvelin K, Beaulieu M, Baeza-Yates R, Hyon Myaeng S (eds) *Proceedings of 25th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2002)*. ACM Press, New York, p 450
- Spence R (2007) *Information visualization: design for interaction*, 2nd edn. Pearson Education Limited, London
- Ware C (2012) *Information visualization - perception for design*, 3rd edn. Morgan Kaufmann Publishers, San Francisco
- Wong PC, Thomas JJ (2004) Visual analytics - guest editors' introduction. *IEEE Comput Graph Appl* 24(5):20–21
- Zhang J (2001) TOFIR: a tool of facilitating information retrieval - introduce a visual retrieval model. *Inf Process Manag* 37(4):639–657
- Zhang J (2008) *Visualization for information retrieval*. Springer, Heidelberg



# Adopting Systematic Evaluation Benchmarks in Operational Settings



Jussi Karlgren

**Abstract** Evaluation of information systems in commercial and industrial settings differs from academic evaluation of methodology in important ways. Those differences have to do with differing organisational priorities between practice and research. Some of those priorities can be adjusted, others must be taken into account, to be able to include evaluation into an operational development pipeline.

## 1 Evaluation in an Operational Setting Differs from an Academic Setting

Some of the differences between operational and academic settings are obvious, some less so. (“Industrial” or “operational” will here be understood to include all kinds of applied uses of information systems, including non-commercial and public contexts of use).

Firstly, an information access service is seldom the primary objective of an industrial project. The industrial project is built to be used for some concrete purpose and information access is a component, frequently an important one, in some process to contribute to that purpose. The ultimate objective of the information access system is to be a sustainable component in that process, for the length of time that process contributes interestingly to the overall goals of that project, be it to generate revenue or goodwill or general happiness.

Secondly, the objective for an industrial project is to perform some task adequately. There is rarely need for optimising performance beyond what is necessary to satisfy the requirements posed on a system. This is in contrast with academic projects, where the goal is to improve and optimise some method, some algorithm, or some performance for some fixed and well specified task. Such improvement and

---

J. Karlgren (✉)  
Gavagai and KTH Royal Institute of Technology, Stockholm, Sweden  
e-mail: [jussi@gavagai.se](mailto:jussi@gavagai.se)

optimisation may not be in the interest of an operational service, in face of limited resources: funds, competent personnel, or attention, all of which are scarce in most industrial contexts.

These two differences have an impact on evaluation methods.<sup>1</sup> What stands in the way of systematic and continuous formal evaluation of information system quality in industry is that evaluation in academic projects focusses on less complex, idealised tasks than what industrial applications or technology can accommodate and evaluation metrics and methods from academic research projects typically reduce an information need challenge into something very clear-cut and clean. Thus, the evaluation schemes proposed in laboratories frequently appear to be irrelevant to understanding the quality of the operational service being offered to customers or end users.

Simplicity and crispness do not reflect the reality of deployed systems in practical use: systems may have many instances, sometimes non-identical; usage may be distributed across numerous nodes; the data under consideration may vary; and the users may have very different objectives than is assumed in an evaluation scheme. Operational data can be messy, incomplete, and distributed over numerous systems, where academic test collections have been cleaned, simplified, and organised to the point that they no longer adequately represent the complexity and variability of the operational realities (Imhof and Braschler 2015). One key factor in making evaluation schemes relevant is to acknowledge the simplification from industrially relevant task to testable output from a system. How then are operational tasks different from those used as models for benchmarking evaluation?

1. The information need may be complex and involve combinations of information items, which makes search technology but one component in a larger whole: “Is this political question worth taking a stand on?” “What factors appear to worry potential customers for our product at what stage in their purchase path?” “What factors in the pension system cause most confusion for our senior citizens?” “Does this group of people pose a risk for public safety?” “Will it be easy or difficult to recruit college graduates to this business area next Fall?”
2. Establishing whether a need is fulfilled or not may be more challenging than in a topical retrieval experiment. The analysis may involve several steps beyond the retrieval or identification of candidate items, and the relevance of such items may be impossible to assess at search time. The determination of what is important, relevant, valuable, or not may be made by someone other than the person who formulates the information need. Sometimes no result is the most positive result, but a *no items found* result page is unsatisfying and not what most analysts hope for. “What published work might be relevant to assessing the novelty of this

---

<sup>1</sup>This point has been made in several recent projects such as CHORUS, TrebleCLEF, or PROMISE, where industrial and research interests have met in think tanks and workshops to share experiences (Braschler 2009; Braschler et al. 2012, e.g.). This insight is also integral to several CLEF evaluation workshops.

- potential patent application?” “Did our customers notice that we mis-labeled the content of our product and corrected it and if they do, do they care?”
3. The real world data and process may be complex and dynamic compared to the analysis of documents from a relatively static benchmarking database. A test set built on a static model may not generalise well. “Do items posted on that video streaming site infringe on our copyright?” “Is the pricing of this tradeable asset moving in some direction?” “How should we set the initial odds for this bet in our book?” “Will the data from our newly acquired division merge well with what we have been working on before?”
  4. The presentation, packaging, and delivery may be complex; the fulfilled information need may not be operational or actionable: even a well executed retrieval or filtering task may not actually deliver what is useful for the organisation. In most organisations, providing more information for decision making means more work, not less, and this may cause some consternation for decision makers at the receiving end. In general, queries such as “What are some of the more interesting trends in our market area that are likely to influence our sales five years down the road?” will provide more useful data than “How many mentions did our brand get in social media and in what sentiment were they expressed?” and in a streaming Big Data access scenario, the individual data points are less interesting than patterns in their flow and changes in those patterns.
  5. In real life tasks, human system users are adaptable and have great readiness to accommodate even to clumsy systems in order to accomplish or further their goals. Applications built from overly simplistic assumptions about user needs may still be functional as tools, and they influence the usage and inform the expectations of users. The cost of introducing new tools, retraining personnel and readjusting processing pipelines may be considerably more complex than coping with noisy or otherwise substandard output from an information system.

## 2 Openness and Accessibility

While academically accepted testing may be attractive for marketing reasons to achieve authority or status, organisations may be skittish to make test results public or use public test sets for reasons related to contractual obligations, commercial risks (real or perceived), or user privacy. If tests are performed in-house, the interpretation of test results may be difficult: if management poses unrealistic goals, which is not unheard of, those in the organisation who are responsible for engineering efforts may be unwilling to provide quantitative data to avoid argumentation and thus unwilling to openly evaluate systems for which they may have less responsibility. And crucially, many academic test sets, if relevant and interesting, are only available to non-profit or research organisations. *A challenge for those who define evaluation schemes and procedures is to make them available for all, and to allow for testing without publication of results.*

### 3 Reliability vs Validity

A method—an algorithm, a computational approach, a memory model etc—may be interesting for research purposes: it may provide insights into human information processing, it may demonstrate interesting characteristics of a collection or the items in it, or it may at some time in the future be the basis for other methods of interest. That method may even score well on various quantitative tests, improving results given by previous approaches. That same method may still be completely uninteresting for practical purposes. A test, however formalised and solid, however robust in its ranking of various experimental conditions, does not guarantee usefulness.

This distinction between *reliability* and *validity* has a long history in the behavioural sciences. Evaluation of information access has for many years been systematic and quantitative, using well-established and commonly accepted benchmarks to compare approaches and methods. These benchmarks, however well normalised and graded, do not guarantee validity of the test. The validity on the test hinges crucially on the task it is patterned to emulate. If the evaluation concerns some behaviour of some component which at the end of the system pipeline makes little or no difference for satisfying the requirements of users, it will have little validity. By contrast, if we want evaluation efforts to predict subsequent take-up of some solution in practice, the evaluation scheme and the metrics it offers need to have high *validity*.

The link between benchmarking a component and assessing its eventual effect on user satisfaction and thus potential for industrial take-up is confounded by a large number of variables, some of which are very challenging to model with any level of confidence in evaluation efforts. If no such linkage can be demonstrated, it is unlikely that the results of an evaluation scheme will convince an industrial system designer to pay attention to that specific evaluation result.

This is where some representation which demonstrates the connection between a system component and its performance on the one hand and user satisfaction on the other will come in handy. In discussions at Conference and Labs of the Evaluation Forum (CLEF), and other related conferences and workshops *use cases* have been proposed as one such potential representation. A use case is a relatively informal or semi-formal description of a system's behaviour and usage intended to capture its functional requirements by describing the interactions between outside agents and the system. Everything should be described in terms with which primary users reach their goals and the description should be useful for system development and evaluation purposes. The objective of using use cases is to make such descriptions simple, lightweight, and incrementally amendable.<sup>2</sup>

---

<sup>2</sup>A use case is *not* a set of scenarios, nor need it be a formal UML schema. Currently, the term use case is often used to mean a vaguely stated area of potential application or a usage scenario for a technology. A use case should be more specific to be useful for system development, and in this case, evaluation (Jacobson 1993).

Use cases for information access evaluation can be written to make hypotheses about user preferences, goals, expectations, and satisfaction explicit. Use cases may be put together with various levels of ambition, competence, and insight. There is no need to aim for perfection, but once formulated, they will enable practitioners and system architects to examine those hypotheses and to assess if an evaluation scheme is relevant to what they are putting effort into and whether it conforms to the behaviour they can observe in their customers and clients. *Use cases (or some similar semi-formal approach) can be used to bridge the gap between benchmarking and validation.*

## 4 The Implicit Use Case of Benchmarking

It is worth noting that the lack of an explicit use case does not mean that there is no use case in the background. The Cranfield paradigm (Cleverdon et al. 1966) compares the capability of information retrieval algorithms to identify and rank topically relevant documents given a well-defined information need under controlled test settings. This, together with appropriate gold standards and scoring practices, has given the information retrieval development efforts a level playing field of immense usefulness. The entire point of that test framework is to abstract evaluation away from variation of factors such as the goal of the user, situation, context, user preferences or characteristics, interaction design, network latency and other such system-external qualities, systematically and intentionally ignoring factors relating to human behaviour and human interaction with information systems. These interaction-related factors will oftentimes be the most important determinants for the user experience of a system, especially if the information retrieval system is but a component in a larger service. *To catch the attention of industrial parties and to ensure validity of their metrics, academic experiments must formulate use cases which capture aspects of interest in deployed tasks.*

## 5 Organisational Thresholds for Introducing Systematic Evaluation in Industrial Projects

The above factors—use case discrepancy, complexity vs measurability, satisficing vs optimisation, lack of resources—all contribute to lack of interest for systematic and routine evaluation of information systems in practical settings, even where it would be motivated. They all contribute to organisational thresholds, which have repeatedly been brought to the fore at discussions in workshops and panels on evaluation in industry (Forner et al. 2013; Kazai et al. 2016; Kanoulas and Karlgren 2017).

Enterprises often lack the resources, above all in terms of engineering personnel, to develop evaluation practice and to keep track of best practice in evaluation research. New graduates who may have performed rigorous evaluation in educational and graduation projects have small possibilities to change existing routines and practices in the organisation they are recruited to work in. Retaining and encouraging the experimental and daring technology culture from the educational background of new entrants is a challenge for any development-oriented organisation, but can if well formulated, have the beneficial side effect to be a persuasive recruitment strategy.

Commercial or related practical realities do not prioritise quality metrics of the type discussed in this volume. Enterprise needs are different from the most generalised needs of the implicit benchmarking use case (Kruschwitz et al. 2017, e.g.). Customers or other end users make multi-factor decisions based on technical and administrative fit to other existing systems and on a multitude of technical factors such as platform independence, scalability, consistency, coverage, and reliability of service, where content quality of output is only one of several features of interest. At the time when a major introduction decision is made, it is likely to be of high priority, but monitoring it continuously fades to the background as the system is installed and deployed. Feedback from end users is handled by customer service and sales staff who have a different focus than engineering staff would. Concrete bug reports will be sent from support staff or sales staff to engineering staff, but more general views of quality of service are routinely covered through workarounds, customer training, or new product releases, the effect of which are more notable for the customer than search component quality. Organisational gaps between customer opinion and engineering staff makes quality monitoring less organisationally useful: using the customer feedback pipeline to motivate continuous quality improvement, not only assurance, will add urgency to quality testing and evaluation. This means turning observations from evaluation metrics into development tickets with concrete goals for improvement of output. *A challenge for industrial and other applied organisations is to encourage a culture of continuous improvement in their technology departments and to provide an information pipeline to support it.*

The focus of a system in production is on its entire output. This is in the end evaluated through sales and customer satisfaction, metrics which have the attention of executive management of an organisation. Component-wise evaluation is done by engineering departments, through systematic testing, most notably through unit testing. Unit testing, the systematic and routine quality testing of components which are subject to development and change, is most often binary in nature: a module passes or fails a test. Quality testing of information retrieval components, by contrast, will yield a score ranging somewhere in the middle between complete failure and perfect ideal performance. The output of such tests is less obviously actionable: an evaluation score from a retrieval test typically does not generate a bug report but may instead invite tuning or improvement efforts. How much effect such an effort has on the bottom line of the organisation can be difficult to assess, and there are no obvious cut-off thresholds that can be set at the outset

to categorise scores into failure vs success. *Industrial sites will need help from academic practitioners to interpret evaluation scores, related to best practice, rather than optimisation.*

In many operational contexts the number of testable components can be prohibitively large. If the engineering effort of a corporation or public office ranges over dozens of different systems many of which have proprietary information access components, some of which are internal to the system, some outward facing, their testing cannot easily be coerced into the same framework. Engineering in a large organisation can be driven by innovation and development efforts as well as maintenance and upkeep. The former efforts involve feasibility decisions and extensive testing; the latter, frequently, assume technology to be stable. This may not always be true, especially in face of changing influx of data and scalability concerns: if the original metrics to motivate a decision have been lost or discarded along the line, reintroducing them will be a challenge and involve a serious amount of work and effort. Evaluation needs to be viewed as part of system monitoring, not solely as a decision making criterion. *Preserving evaluation metrics from development processes and keeping them in place during the operation life cycle phase of a system saves effort.*

## **6 How to Make Evaluation Practice Relevant for Industry**

The main lessons to be learnt from examining the gap between academic and operational evaluation are that to make the former more relevant and the latter more systematic and actionable, the operational priorities of a system development process need to be taken into account and adjusted where necessary.

- Evaluation schemes and procedures must be conveniently available and integrable to allow for testing without publication of results.
- Evaluation target notions, methods, procedures, and metrics must have validity with respect to tasks.
- Validity can be achieved through e.g. formulation of use cases which capture aspects of interest in deployed tasks.
- Evaluation schemes must be sensitive to the distinction between optimisation and best practice.
- Many evaluation schemes, while useful benchmarks for academic research, will not be useful for industrial sites.
- Industrial sites will need help from academic practitioners to interpret evaluation scores.
- Industrial organisations must recognise that development and deployment decisions feed into the entire life cycle of a system.
- Industrial organisations must encourage a culture of continuous improvement.
- Industrial organisations must provide an information pipeline and procedures to support such a culture.

## References

- Braschler M (2009) Best practices in system-oriented aspects for multilingual information access applications. In: Proceedings of the eChallenges 2009 conference
- Braschler M, Rietberger S, Imhof M, Järvelin A, Hansen P, Lupu M, Gäde M, Berendsen R, de Herrera AGS (2012) Best Practices Report, Deliverable 2.3. PROMISE project
- Cleverdon CW, Mills J, Keen M (1966) Aslib Cranfield research project—factors determining the performance of indexing systems. Technical report
- Forner P, Bentivogli L, Braschler M, Choukri K, Ferro N, Hanbury A, Karlgren J, Müller H (2013) PROMISE technology transfer day: spreading the word on information access evaluation at an industrial event. *SIGIR Forum* 47(1):53–58
- Imhof M, Braschler M (2015) Are test collections “Real”? Mirroring real-world complexity in IR test collections. In: Mothe J, Savoy J, Kamps J, Pinel-Sauvagnat K, Jones GJF, SanJuan E, Cappellato L, Ferro N (eds) *Experimental IR meets multilinguality, multimodality, and interaction. Proceedings of the sixth international conference of the CLEF association (CLEF 2015)*. Lecture notes in computer science (LNCS), vol 9283. Springer, Heidelberg, pp 241–247
- Jacobson I (1993) *Object-oriented software engineering: a use case driven approach*. Pearson Education India, Delhi
- Kanoulas E, Karlgren J (2017) Practical issues in information access system evaluation. *SIGIR Forum* 51(1):67–72
- Kazai G, Ingersoll G, Lin J (2016) Evaluation is for conference papers. I need to build a real life product! *SIGIR 2016 Industry Track Panel*, Pisa
- Kruschwitz U, Hull C et al (2017) Searching the enterprise. *Found Trends Inf Retr* 11(1):1–142



# Author Index

## A

Agosti, Maristella, 105  
Amigó, Enrique, 487

## B

Balog, Krisztian, 511  
Bogers, Toine, 415  
Bonnet, Pierre, 389  
Braschler, Martin, 177

## C

Caputo, Barbara, 275  
Carrillo-de-Albornoz, Jorge, 487  
Clough, Paul, 217

## D

Daelemans, Walter, 461  
Dang-Nguyen, Duc-Tien, 275  
Di Nunzio, Giorgio Maria, 105

## F

Ferro, Nicola, 3, 105, 565  
Forner, Pamela, 441  
Fuhr, Norbert, 555

## G

García Seco de Herrera, Alba, 231  
Geva, Shlomo, 415  
Giampiccolo, Danilo, 441  
Gilbert, Andrew, 251

Glotin, Hervé, 389  
Goeuriot, Lorraine, 333  
Gollub, Tim, 123  
Gonzalo, Julio, 487  
Goëau, Hervé, 389

## H

Halvorsen, Pål, 275  
Hanbury, Allan, 161, 365  
Hedlund, Turid, 201  
Hopfgartner, Frank, 511  
Hovy, Eduard, 441

## J

Järvelin, Kalervo, 201  
Joly, Alexis, 389  
Jones, Gareth J. F., 307

## K

Kalpathy-Cramer, Jayashree, 231  
Kamps, Jaap, 415  
Karlgren, Jussi, 201, 583  
Kelly, Liadh, 333, 511  
Keskustalo, Heikki, 201  
Kettunen, Kimmo, 201  
Kille, Benjamin, 511  
Koolen, Marijn, 415

## L

Larsen, Birger, 547  
Larson, Martha, 511

Lombardo, Jean-Christophe, 389  
Lommatzsch, Andreas, 511

**M**

Magnini, Bernardo, 441  
Müller, Henning, 161, 231, 389

**P**

Palazzo, Simone, 389  
Peters, Carol, 3  
Peñas, Anselmo, 441  
Piras, Luca, 275  
Piroi, Florina, 365  
Planqué, Robert, 389  
Potthast, Martin, 123, 461

**R**

Rangel, Francisco, 461  
Riegler, Michael, 275  
Rodrigo, Álvaro, 441  
Rosso, Paolo, 461

**S**

Sakai, Tetsuya, 71

SanJuan, Eric, 415  
Santucci, Giuseppe, 565  
Savoy, Jacques, 177  
Schenkel, Ralf, 415  
Schuth, Anne, 511  
Silvello, Gianmaria, 105  
Spampinato, Concetto, 389  
Stamatatos, Efstathios, 461  
Stein, Benno, 123, 461  
Suominen, Hanna, 333  
Sutcliffe, Richard, 441

**T**

Thomee, Bart, 251  
Tsikrika, Theodora, 217

**V**

Vellinga, Willem-Pier, 389  
Villegas, Mauricio, 251  
Voorhees, Ellen M., 45

**W**

Wang, Josiah, 251  
Wiegmann, Matti, 123

# Subject Index

- Ad-hoc retrieval, 11, 18, 27, 49, 56, 178, 418  
Anaphor, 208  
Answer validation, 16, 24, 59, 448  
Author identification, 26, 157, 462, 469, 481  
Author profiling, 26, 462, 474, 481, 492, 494, 505
- Benchmarking, 6, 10, 162, 167, 232, 268, 269, 282, 299, 390, 462, 480, 525, 583  
Best practice, 391  
Bibliometrics, 39, 112, 547, 548, 553  
Bilingual information retrieval, 184  
Biodiversity information retrieval, 29, 390  
Biomedical retrieval, 29, 231  
Birdclef challenge, 396  
Blind evaluation, 462
- Caption prediction, 233, 236, 239  
Citation analysis, 39, 377, 550  
Citation impact, 552, 553  
Classification, 264, 371, 374, 390, 430, 431, 448, 472, 480, 500, 502, 523  
Clef ehealth, 29, 335, 336, 341–343, 349  
Clef-ip, 26, 365  
Cloud-based evaluation, 129, 131, 134, 159, 167, 168, 462, 467, 481  
Clustering, 280, 285, 287, 297, 462, 468, 470, 499  
Collaborative search, 31  
Commercial settings, evaluation in, 583  
Complex information needs, 32  
Complex search task, 32
- Compound terms, 202, 206  
Computer vision, 260, 268, 276, 295  
Conceptual model, 110  
Content-based retrieval, 231, 276, 285, 286, 390, 520, 521  
Contextual suggestion track, 421, 422, 424, 427, 428  
Convolutional neural network, 243, 260, 264, 266, 269, 291–293, 295, 297, 395, 400, 480  
Cranfield paradigm, 6, 45–48, 51, 53, 56, 57, 65, 106, 110, 161, 512  
Cross-language information retrieval, 12, 56, 178, 201, 219, 222, 375  
Crowd artificial intelligence, crowd AI, 225, 244  
Cultural differences, 32, 56, 204  
Cultural microblog contextualization, 32, 429–432  
Cyber-security, 461
- Data fusion, 243  
Data management, 567  
Deep learning, 243, 266, 267, 270, 277, 292, 293, 297–299, 395, 400, 480  
Dictionary based cross-language information retrieval, 185, 211  
Digital library, 27, 209, 417, 455  
Digital text forensics, 26, 128, 462, 480  
DIRECT, 106, 109, 114, 116, 118, 561  
Diversity, 220, 221, 254, 282, 283, 285–287, 296  
Document translation, 194

- Domain-specific retrieval, 13, 275
- Electronic patient records, 29
- Evaluation-as-a-service, 40, 129, 156, 159, 161, 164, 462, 467, 481, 513, 537, 560
- Evaluation infrastructure, 4, 40, 45, 109–111, 113, 119, 163, 462, 467, 481, 513, 524
- Evaluation initiative, 8, 45, 109, 269, 275, 462, 506, 516, 526
- Evaluation lab, 21, 252, 275, 365, 368, 390, 462, 480, 516, 526
- Evaluation measure, 92
- Evaluation workflow, 109, 110, 390, 571, 573, 580
- Experimental data, 106, 111, 112, 119
- Experimental evaluation, 6, 106, 109, 117, 118, 218, 390
- Factoid question, 58, 441
- Failure analysis, 571
- Filtering task, 18, 57, 430, 431, 498, 521
- FIRE, 9, 46, 56, 108, 462, 555
- Geographic retrieval, 18
- Image annotation, 16, 234, 236, 240, 252–254, 256, 257, 260, 261, 263, 264, 267, 268, 270, 276, 282, 287, 288, 290, 291, 293
- Image captioning, 260–262, 267, 269, 270
- Image classification, 16, 264, 268, 279–282, 288, 374, 392
- ImageCLEF, 16, 25, 218, 231, 252, 253, 257, 264, 268–270, 275–277, 282, 287, 293, 335
- ImageCLEF medical retrieval, 231, 335
- Image retrieval, 16, 25, 217, 231, 252, 275–277, 282, 286, 292, 297, 298
- Impact of conferences, 548
- Indexing, 454
- INEX, 10, 27, 58, 59, 109, 378, 416–425
- Inflection, 202, 204
- Information extraction, 22–24, 33, 335, 338, 340, 341, 450
- Information need, 6, 46, 47, 53, 57, 115, 422, 426, 428, 454, 583
- Information visualization, 336, 338, 339, 565, 579
- Interactive retrieval, 15, 31, 56, 298, 423, 425, 427, 538, 560
- Interactive social book search, 31, 423, 425, 427
- Inter-assessor agreement, 50, 85, 477
- Jump-in point, 311
- Language typology, 201
- Language variation, 53, 56, 179, 181, 201, 431, 474
- Lifelogging, 252, 275, 293
- Linguistic typology, 201
- Linked open data, 106, 113, 116, 117
- Living lab, 30, 513–515, 518, 526, 560
- Low-density language, 201
- Mean average precision (MAP), 7, 51, 53, 54, 61, 63, 64, 239, 255, 259, 265, 283, 381, 398, 491, 558
- Mean generalised average precision (mGAP), 318
- MediaEval, 321
- Medical image retrieval, 25, 231
- Metadata, 12, 27, 112, 253, 255, 256, 282, 285–287, 407
- Microblog search, 32, 33, 420, 422, 423, 429, 430
- Modality classification, 24, 236
- Morphological variation, 181
- Morphology, 35, 202, 204
- Multi label classification, 277, 288
- Multilingual information retrieval, 4, 11, 12, 18, 27, 190, 201, 218
- Multilingual question answering, 16, 24, 442
- Multimodal information retrieval, 4, 20, 218, 232, 320, 392
- Named entity linking, 24
- Natural language processing (NLP), 12, 16, 201, 297, 334–336, 427, 431, 442, 462, 488, 506
- News recommender system, 30, 516
- News retrieval, 12, 30
- NTCIR, 9, 46, 56, 108, 294, 295, 323, 368, 446, 555
- Object recognition, 252, 268, 270, 279, 281, 392

- Offline evaluation, 6, 518  
 Online evaluation, 15, 30, 32, 512, 517, 526  
 Online reputation management, 28, 475, 488  
 Ontology, 18, 255, 258, 260, 451  
 OpenCLIR, 325
- PAN, 26, 128, 153, 159, 462, 480  
 Patent office, 367  
 Patent retrieval, 26, 368  
 Personalized retrieval, 33  
 Pivot language, 194  
 Plagiarism detection, 26, 156, 462, 463  
 Plant identification, 29, 252, 392  
 Pooling, 48, 49, 56, 58, 60, 63, 82, 559, 575  
 Precision, 7, 47, 50, 58, 62, 63, 267, 283, 290, 296, 446  
 Prior art, 26, 369, 370, 374  
 Product search, 526
- Query translation, 188  
 Question answering, 16, 24, 58, 442
- Ranking, 6, 46, 57, 62, 64, 65, 285, 293, 390, 456, 489–491, 500  
 Recall, 7, 47, 50, 51, 58, 62, 266, 267, 283, 290, 446  
 Recommender system, 30, 426–428, 515  
 Relevance assessment, 6, 46, 48, 49, 52, 56, 61, 62, 84, 234, 237, 316, 377, 538, 577  
 Relevance feedback, 55, 57, 283, 420  
 Reproducibility, 10, 34, 41, 108, 113, 123, 126, 129, 159, 163, 169, 391, 466, 467, 481, 513, 518, 539, 559  
 Reputational polarity, 28, 492, 498, 504  
 Retrieval effectiveness, 6, 45, 47, 50, 55, 512, 567  
 Reusability, 79  
 Robot vision, 275, 277  
 Round robin, 190
- Scholarly impact of CLEF, 548  
 Scientific data, 111, 113, 119, 169, 567  
 Search result diversification, 96  
 Semantic annotation, 252, 420–422, 424
- Sentiment analysis, 498  
 Shared task platform, 124, 125, 131, 157, 159, 462, 467, 480  
 Snippet retrieval, 420, 422  
 Social book search, 31, 425–429  
 Species identification, 29, 390  
 Speech recognition, 17, 309, 338, 340  
 Speech retrieval, 308  
 Spoken content retrieval, 316  
 Spoken document retrieval, 312  
 Statistical significance, 8, 55, 92, 96, 558, 577  
 Statistical translation model, 258, 262  
 Stemming, 181  
 Stopword list, 179  
 Style breach detection, 462, 471
- Task design, 7, 57, 374, 488  
 Test collection, 6, 35, 45–50, 53, 55, 59, 62, 107, 372, 464, 473, 475, 500  
 Text alignment, 157, 464  
 TIRA, 118, 124, 128, 133, 151, 462, 467  
 Topic analysis, 111  
 Translation strategy, 185  
 TREC, 8, 45, 46, 48–50, 53, 55, 58, 107, 128, 335, 368, 421, 422, 424, 427, 428, 442, 466, 514, 555, 559  
 TREC Spoken Document Retrieval (SDR), 312  
 TRECVID, 313, 318, 322  
 Tweet contextualization, 32
- Use case, 390, 516, 529, 534, 583
- Validity, 49, 51, 370, 557, 586  
 Vandalism detection task, 154, 462  
 Video retrieval, 308  
 Virtual machine, 128, 138, 156, 165, 166, 525  
 Visual analytics, 20, 41, 112, 566, 579  
 Visual feature, 234, 256, 258, 265, 282, 283, 286, 291, 297, 392  
 Visual information retrieval, 217, 231, 282, 308, 392
- Wikipedia, 222, 260, 320, 441, 462  
 Word decompounding, 183