

Giuseppe Serra
Carlo Tasso (Eds.)

Communications in Computer and Information Science

806

Digital Libraries and Multimedia Archives

14th Italian Research Conference on Digital Libraries, IRCDL 2018
Udine, Italy, January 25–26, 2018
Proceedings

 Springer

EXTRAS ONLINE

Communications in Computer and Information Science

806

Commenced Publication in 2007

Founding and Former Series Editors:

Alfredo Cuzzocrea, Xiaoyong Du, Orhun Kara, Ting Liu, Dominik Ślęzak,
and Xiaokang Yang

Editorial Board

Simone Diniz Junqueira Barbosa

*Pontifical Catholic University of Rio de Janeiro (PUC-Rio),
Rio de Janeiro, Brazil*

Phoebe Chen

La Trobe University, Melbourne, Australia

Joaquim Filipe

Polytechnic Institute of Setúbal, Setúbal, Portugal

Igor Kotenko

*St. Petersburg Institute for Informatics and Automation of the Russian
Academy of Sciences, St. Petersburg, Russia*

Krishna M. Sivalingam

Indian Institute of Technology Madras, Chennai, India

Takashi Washio

Osaka University, Osaka, Japan

Junsong Yuan

Nanyang Technological University, Singapore, Singapore

Lizhu Zhou

Tsinghua University, Beijing, China


More information about this series at <http://www.springer.com/series/7899>


Giuseppe Serra · Carlo Tasso (Eds.)

Digital Libraries and Multimedia Archives

14th Italian Research Conference on Digital Libraries, IRCDL 2018
Udine, Italy, January 25–26, 2018
Proceedings

Editors

Giuseppe Serra 
University of Udine
Udine
Italy

Carlo Tasso 
University of Udine
Udine
Italy

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-319-73164-3 ISBN 978-3-319-73165-0 (eBook)
<https://doi.org/10.1007/978-3-319-73165-0>

Library of Congress Control Number: 2017962896

© Springer International Publishing AG 2018

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.

Printed on acid-free paper

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

Preface

In 2018 the Italian Research Conference on Digital Libraries reached its 14th edition. Since 2005, it has served as an important national forum focused on digital libraries and related technical, practical, and social issues. IRCDL encompasses the various facets of the term “digital libraries,” including: new forms of information institutions; operational information systems with multimodal digital content; new means of selecting, collecting, organizing, and distributing digital content; and theoretical models of information media, including document genres and electronic publishing. The conference continues, year after year, to address new themes and challenges that witness the long-lasting evolution and impact of digital libraries.

Representatives from academia, government, industry, and others are invited to participate in this annual conference. The conference draws from a broad and multidisciplinary array of research areas including computer science, information science, librarianship, archival science and practice, museum studies and practice, technology, social sciences, and humanities. The national Program Committee comprised 32 members, with representatives of the most active Italian research groups on digital libraries.

This volume contains the accepted papers at the 14th Italian Research Conference on Digital Libraries (IRCDL 2018), which was held at the Palazzo di Toppo Wassermann in Udine (Italy) during January 25–26, 2018. Starting from 30 submissions, after receiving three reviews per paper, 25 papers were presented to the conference and were accepted to be published in this volume. The covered topics are related to the different aspects of digital libraries:

- Multimedia digital libraries
- Formal and methodological foundations of digital libraries
- Digital libraries architectures and infrastructures
- Algorithms and techniques for automatic content analysis
- Text analysis and mining
- Advanced multimodal access to digital libraries
- System interoperability and data integration
- User interfaces and visualization
- Information access, usability, and personalization
- Exploitation of digital cultural heritage collections
- New models and applications for digital exhibitions
- Current and relevant projects in the field of digital libraries

The 2018 edition of IRCDL was organized with the patronage and the sponsorship of the Laboratory of Artificial Intelligence of the University of Udine.

We would like to thank all the authors for their contributions. In particular, we would like to thank the Program Committee members and the Steering Committee members who took part in the evaluation of manuscripts, and provided insights that

resulted in a comprehensive volume. The program chairs would also like to thank the Department of Scienze Matematiche, Informatiche e Fisiche of the University of Udine and Springer for the exceptional support and effort they provided throughout the entire process.

January 2018

Giuseppe Serra
Carlo Tasso

Organization

Program Chairs

Carlo Tasso
Giuseppe Serra

University of Udine, Italy
University of Udine, Italy

Steering Committee

Maristella Agosti
Tiziana Catarci
Alberto Del Bimbo
Floriana Esposito
Carlo Tasso
Costantino Thanos

University of Padua, Italy
University of Rome La Sapienza, Italy
University of Florence, Italy
University of Bari, Italy
University of Udine, Italy
ISTI CNR, Pisa, Italy

Program Committee

Giovanni Adorni
Elisa Antolli
Lamberto Ballan
Lorenzo Baraldi
Valentina Bartalesi Lenzi
Marco Basaldella
Marco Bertini
Maria Teresa Biagetti
Simone Calderara
Diego Calvanese
Vittore Casarosa
Michelangelo Ceci
Fabio Ciotti
Marcella Cornia
Rita Cucchiara
Stefano Ferilli
Nicola Ferro
Costantino Grana
Maria Guercio
Donato Malerba
Paolo Manghi
Simone Marinai
Stefano Mizzaro
Nicola Orio
Antonella Poggi

University of Genoa, Italy
University of Udine, Italy
University of Padua, Italy
University of Modena and Reggio Emilia, Italy
ISTI-CNR, Italy
University of Udine, Italy
University of Florence, Italy
University of Rome La Sapienza, Italy
University of Modena and Reggio Emilia, Italy
Free University of Bozen-Bolzano, Italy
ISTI-CNR, Italy
University of Bari, Italy
University of Rome Tor Vergata, Italy
University of Modena and Reggio Emilia, Italy
University of Modena and Reggio Emilia, Italy
University of Bari, Italy
University of Padua, Italy
University of Modena and Reggio Emilia, Italy
University of Rome La Sapienza, Italy
University of Bari, Italy
ISTI-CNR, Italy
University of Florence, Italy
University of Udine, Italy
University of Padua, Italy
University of Rome La Sapienza, Italy

Marco Schaerf	University of Rome La Sapienza, Italy
Lorenzo Seidenari	University of Florence, Italy
Luigi Siciliano	University of Bolzano, Italy
Gianmaria Silvello	University of Padua, Italy
Francesca Tomasi	University of Bologna, Italy
Fabio Vitali	University of Bologna, Italy
Paul Gabriele Weston	University of Pavia, Italy

Contents

Digital Library Architecture

Subject Access to Images and Exploratory Search	3
<i>Andrea Cuna</i>	
Library Data Integration: The CoBiS Linked Open Data Project and Portal	15
<i>Luisa Schiavone, Federico Morando, Davide Allavena, and Giorgio Bevilacqua</i>	
A Software Architecture for Narratives	23
<i>Carlo Meghini, Valentina Bartalesi, Daniele Metilli, and Filippo Benedetti</i>	
Thirty Years of Digital Libraries Research at the University of Padua: The Systems Side	30
<i>Maristella Agosti, Giorgio Maria Di Nunzio, Nicola Ferro, and Gianmaria Silvello</i>	
Thirty Years of Digital Libraries Research at the University of Padua: The User Side	42
<i>Maristella Agosti, Giorgio Maria Di Nunzio, Nicola Ferro, Maria Maistro, Stefano Marchesin, Nicola Orio, Chiara Ponchia, and Gianmaria Silvello</i>	

Multimedia Content Analysis

An Abstract Argumentation-Based Approach to Automatic Extractive Text Summarization	57
<i>Stefano Ferilli and Andrea Pazienza</i>	
On Frequency-Based Approaches to Learning Stopwords and the Reliability of Existing Resources — A Study on Italian Language	69
<i>Stefano Ferilli and Floriana Esposito</i>	
<i>Text</i> - Text Extractor Tool for Handwritten Document Transcription and Annotation	81
<i>Anders Hast, Per Cullhed, and Ekta Vats</i>	
The Distiller Framework: Current State and Future Challenges	93
<i>Marco Basaldella, Giuseppe Serra, and Carlo Tasso</i>	

Applications of Duplicate Detection in Music Archives: From Metadata Comparison to Storage Optimisation: The Case of the Belgian Royal Museum for Central Africa.	101
<i>Joren Six, Federica Bressan, and Marc Leman</i>	
Extracting Dependency Relations from Digital Learning Content.	114
<i>Giovanni Adorni, Felice Dell’Orletta, Frosina Koceva, Ilaria Torre, and Giulia Venturi</i>	
Annote: A Serious Game for Medical Students to Approach Lesion Skin Images of a Digital Library	120
<i>Fabrizio Balducci</i>	
Term-Based Approach for Linking Digital News Stories	127
<i>Muzammil Khan, Arif Ur Rahman, and Muhammad Daud Awan</i>	
A Graphic Matching Process for Searching and Retrieving Information in Digital Libraries of Manuscripts	139
<i>Nicola Barbuti, Tommaso Caldarola, and Stefano Ferilli</i>	
XDOCS: An Application to Index Historical Documents	151
<i>Federico Bolelli, Guido Borghi, and Costantino Grana</i>	
Object Recognition and Tracking for Smart Audio Guides	163
<i>Lorenzo Seidenari, Claudio Baccchi, Tiberio Uricchio, Andrea Ferracani, Marco Bertini, and Alberto Del Bimbo</i>	
Automatic Image Cropping and Selection Using Saliency: An Application to Historical Manuscripts	169
<i>Marcella Cornia, Stefano Pini, Lorenzo Baraldi, and Rita Cucchiara</i>	
Bidirectional LSTM Recurrent Neural Network for Keyphrase Extraction.	180
<i>Marco Basaldella, Elisa Antolli, Giuseppe Serra, and Carlo Tasso</i>	
Models and Applications	
Eliciting the Ancient Geography from a Digital Library of Latin Texts	191
<i>Maurizio Lana and Timothy Tambassi</i>	
A Research Tool for the ERC-Funded EMOBookTrade Project	201
<i>Giliola Barbero and Luigi Tassarolo</i>	
The Biographical Dictionary of Friulians - “Nuovo Liruti” Online: A Biographical Dictionary Based on Semantic Web and Linked Open Data.	209
<i>Stefano Allegrezza and Nicola R. Di Matteo</i>	

ISS Project: The Integrated Search System in the National
Bibliographic Services 219
Luigi Cerullo

User Requirements and Relational Modelling for a Non-theatrical Cinema
and Video-Art Cataloguing System 225
Petra Marlazzi, Lisa Parolo, Cosetta Saba, and Nicola Vitacolonna

The European Project OpenUP: OPENing UP New Methods, Indicators
and Tools for Peer Review, Impact Measurement and Dissemination
of Research Results 240
Alessia Bardi, Vittore Casarosa, and Paolo Manghi

Who Is the Data Curator? Defining a Vocabulary 249
Anna Maria Tammaro and Vittore Casarosa

Author Index 257

Digital Library Architecture

Subject Access to Images and Exploratory Search

Andrea Cuna^(✉) 

University of Udine, Udine, Italy
andrea.cuna@uniud.it

Abstract. As traces of social life and material culture of the past, non-art images are carriers and prompters of memory. They are important sources for social and cultural history and, at the same time, valuable cultural heritage resources. Cultural heritage information systems (CHISs) very often rely on basic search and browsing features to provide access to information related to non-art digital images. However, these forms of access are not very helpful for non-expert or casual users, who usually move through the information space in an exploratory way. Although significant strides have been made to understand exploratory search activities, there are still some open issues when it comes to the user interface (UI). After briefly reviewing concept-based indexing techniques applied to images, this paper explores some of the questions related to UI design and provides insights into how to develop a browse-and-search framework to enhance exploratory search tasks.

1 Introduction

Like words, images convey meaning, but in a way that has nothing to do with the written or verbal language. Indeed, as the old adage says, a picture is worth a thousand words. Yet, in today's online environment, most indexing is precisely carried out using words, and search engines are predominantly keyword-based. What is more, there may be differences between the indexing system and the search system, resulting in the so-called vocabulary problem and a semantic gap [14].

Most information nowadays available on the Web is non-indexed. As a result, the only feature available to users to find out what they are looking for is free-text searching. This happens through a simple search input box, usually placed at the top right of the page. Here users can enter their queries using keywords. As a response, the system returns a set of matching results, usually sorted by relevance according to some similarity measure. This access pattern is the primary mode of searching offered by cultural heritage information systems (CHISs) and underlies a search model that targets users who have specific information needs in mind and well-defined search tasks to perform [18].

Since users are accustomed to Web search engines and their performance in returning results, retrieval latency plays an important role in the users' satisfaction: the higher the latency, the lower the quality of the results perceived by the

users [12]. However, there are many different kinds of search, ranging from simple lookup tasks, such as known-item search in bibliographic databases, to informal browsing. When users have ill-defined, fuzzy or broad information needs, their behavior falls on the right side of the range. Users engage then in a particular type of information seeking process known as exploratory search (ES).

That ES implies learning and investigation tasks has long been implicitly recognized in the literature: “the query is satisfied not by a single final set, but a series of selections of individual references and bits of information at each stage of the ever-modifying search. A bit-at-a-time retrieval of this sort is here called berrypicking” [5] (p. 409). The berrypicking model aptly describes the dynamism and evolving nature of both search focus and information needs. Its non-linear and segmented path is a good graphical representation of how each piece of new information encountered along the way can result in new ideas and directions to follow. A berrypicking search seems to be an almost informal way of seeking information, albeit not restricted to informal methods only.

Rather than being designed according to the traditional direct search paradigm, CHISs should take a more user-oriented approach and implement navigation models that are capable of supporting those users who find unnatural to begin a search using keywords. These systems should also support those users who employ informal methods, by offering browsing functionalities going far beyond those based on the organization of the collection or context metadata, which do not reflect the image subject(s).

A complementary approach would be to offer a multidimensional browsing space where the subject terms describing what the image is *of* (description) and *about* (identification) are organized according to their paradigmatic relationships. Berrypicking is perhaps the most common strategy that users follow in ESs. Recent research has focused on the features required to best support ES patterns: White and Roth [32] have outlined the challenges of developing an ES system to support ES activities. Briefly, the set of features required for such a system includes: (1) an overview of the information space covered by the system, (2) meaningful representations of results, (3) support for query formulation and refinement, (4) faceted browsing and filtering, and (5) track-keeping of search path and history.

In recent years, faceted search (FS) has emerged as a key feature for enhancing the user search experience. Combining keyword search with faceted navigation has proven very effective as it lets users begin with a simple query and then refine search results by facets [16, 28, 33]. FS has also turned out to be a useful model to support ESs [32], although a few changes are still needed to make it more compliant with ES requirements.

The remainder of this paper is organized as follows: Sect. 2 gives an overview of concept-based indexing techniques applied to images; Sect. 3 focuses on how the concept of facet has been variously interpreted in different domains; Sect. 4 takes a look at hierarchical and faceted systems; Sect. 5 deals with search interaction patterns; Sect. 6 discusses the key characteristics of an exploratory tool for non-expert users; finally, Sect. 7 closes the paper and touches on future work.

2 Subject Access to Images

Photographic images have been the subject of attention of various disciplines, both old and new. Not only have these disciplines played a major role in increasing public awareness of the great social, historical and cultural significance embodied in such “documents of reality”, but they have also developed specific categories and interpretative tools to understand the language of that specific means of expression (Fig. 1).



Fig. 1. Embroiderers (group portrait), Cormóns (GO, Italy) 1899. (Gorizia Archdiocese [Digital] Photographic Archive, C00034).

Over the last few years, libraries, archives and museums (LAMs) have tried to support non-expert users by providing innovative ways of searching and browsing that go beyond the query-response paradigm. Content-based indexing has offered in this regard an alternative way of exploring and engaging with digital cultural heritage resources [27]. At its most basic level, content-based indexing allows users to search for images by means of visual features such as color, shape and texture. However, these are low-level features, with limited descriptive power as regards the semantics of an image. A more suitable approach to improve retrieval is suggested by research conducted in domains like medical informatics, wherein visual content is combined with text-based data [2].

Unlike content-based indexing, concept-based indexing represents *subject matter* through descriptive terms taken directly from natural language, as is the case of social tagging, or through corresponding descriptors chosen from a controlled vocabulary. The latter is a far more formal way of representing topicality and involves various aspects of standardization that are of pivotal importance for CHISs. Vocabulary control is one of the key steps of this process as it is only by using a common terminology that any potential semantic gap existing between indexers and users is bridged and retrieval achieved.

In the case of library systems, concept-based indexing is usually carried out by human indexers. It begins with determining what a document is *about* and

goes on with a conceptual analysis and translation into a controlled terminology. However, it must be said that what *about-ness* really is remains controversial. From a very pragmatic point of view, about-ness can be considered as a variable, mainly affected by the type of language used: referential language refers to the objective reality, hence sentences are about their subjects, while representational or emotive language communicates or excites feelings [26] (pp. 46–48).

While determining the subject of a text according to the so-called grammar model (a careful examination of virtually any sentences making up the text) is somewhat problematic, determining the subject of, say, a photographic image is in general relatively straightforward as its referent or its genuine essence is “la chose nécessairement réelle qui à été placée devant l’objectif, faute de quoi il n’y aurait pas de photographie” [4] (p. 120). However, the referent is not enough to grasp the whole meaning of a photographic image. In Henry Ziegler’s portrait of his cousin Gaspard, the pocket watch shown by the sitter might now suggest a simple reflection on the flow of time, while for a contemporary observer it meant the long time of the pose typical of early photographic processes [6]. This intended meaning is, in Barthes’ words, *studium* [4] (pp. 47–48), which typically requires skills and knowledge other than those possessed by the layperson. Another much more problematic concept coined by Barthes is *punctum*. This is something present in the image that immediately jumps out at the viewers as unusual or unexpected and triggers in them a feeling, thought or memory. Therefore, *punctum* belongs entirely to the sphere of the beholder’s subjectivity [4] (pp. 48–49). Barthes’ considerations seem to offer a good basis for reconsidering the most common concept-based approaches currently underlying subject indexing of images as they provide for a differentiation between intersubjective and subjective meaning.

A complementary, albeit different, approach reported in the vast literature on image indexing is that of German art historian Erwin Panofsky [22]. His *Studies in Iconology* (1939) has had a strong influence on art history research and proven to be a valuable method for analyzing the subject of a work of art. In his model, Panofsky identifies three layers of subject matter in a work of art: (1) primary or natural subject matter (pre-iconographic description); (2) secondary or conventional subject matter (iconographic identification), and (3) intrinsic meaning (iconological interpretation). At the first level, perception yields factual meaning, i.e. referential meaning (objects, events), but psychological nuances (expressional meaning) can possibly emerge as well. At the second level, referents are related to the relevant frame of conventional values (social, art historical etc.). Finally, the third level is concerned with the essential or intrinsic meaning of an image, which is grasped by widening the analysis to the religious, social, historical and philosophical context in which the work of art was created. At this level, a work of art is understood to be symptomatic of a worldview (*Weltanschauung*) [22] (pp. 5–8 and 14–25).

Panofsky’s model has been largely deployed in image indexing, albeit modified to merely fit instrumental and pragmatic needs. Indeed, what was originally conceived as a unitary process geared towards understanding the inner

essence of a work of art that encompasses the whole continuum of the interpretive spectrum has been regrettably turned into no more than a set of procedural guidelines mainly focused on the pre-iconographic and iconographic levels (see for example [19]). For this, as well as other reasons, Panofsky's model (1939) has been recently criticized and challenged by some scholars, who have called for a rethinking of the theoretical and practical foundations of image indexing (for example [10]). However, when going back to the apparently banal example given by Panofsky in the *Introductory* of his *Studies* (1939) – a natural event of the everyday life – one cannot overlook the fact that it was purposely chosen “to exemplify the **minimal** [bold mine] features of visual communication and representation”, and to serve as a “baseline from which to measure more complex” forms of “visual representation” (cf. [21], p. 26). Therefore, it is not a matter of doing Panofsky's model anew or simply to scale it down from the level of iconological interpretation to that of the minimal features of non-art imagery. Rather one should acknowledge that there are methodological differences in approaching the problem of subject indexing of images and that the building blocks of a heuristic model for subject access to non-art imagery are to be laid on the basis of the analysis of images that represent events of everyday life like the one described in Panofsky's model.

Shatford Layne's contribution to image indexing, on both the theoretical and practical level, appears to address many of the issues surrounding this point [25]. Her framework for analyzing the subject of an image consists of: (1) a description in generic terms (*generic of-ness*), (2) identification in specific terms (*specific of-ness*), and (3) interpretation (*about-ness*). The main limit of this framework becomes immediately clear when analyzing Fig. 1. The content of this photographic image is actually readable only at the first two levels: (1) women and children (*generic of-ness*), and (2) embroiderers (*specific of-ness*). The third level of meaning, i.e. the meaning to which the image referent alludes, is problematic because there is no readily available information in the image pointing to some underlying message. If one wants to find out such message, they have to rely upon the help of sources other than the image itself, which may not be readily available, or even exist at all.

As far as art images are concerned, *about-ness* is core to subject analysis, especially when symbols and allegories are at work, or when it is stated or clearly apparent. However, when this is tenuous and ambiguous, because the interpretation is heavily dependent on the observer's subjectivity, it can be even omitted altogether. *About-ness* then is the constitutive factor that differentiates art images from non-art images. However, since non-art images are not only a representation of reality but also, in Barthes' words, prompters of *punctum*, it follows that: (1) the analysis of non-art images should not be limited to the generic-to-specific continuum of the spectrum making up the description and identification of the reality depicted in an image in order to fit it into categories according to some schema; (2) it is appropriate to provide additional or alternative points of view, i.e. access points to the who, what, when and where that are right **there** (*there-ness*) **in**, rather than **about**, an image. If neither the interaction between

studium and *punctum* nor the rules for knowing the *punctum* can be established in advance, then the analysis of the content of non-art images should take place solely at the of-ness level. And the related subject metadata created by indexers should adequately reflect the outcome of this analysis at the level of detail as required for the so-called minimal features of non-art images.

As regards this issue, it was considered appropriate to conform to the version of *Subject Matter* presented in [15], comprising the following sub-categories: *description*, *identification*, and *interpretation*. Additionally, although [15] provides everything necessary to develop a search interface, the way in which the descriptive content was structured for the project in question is quite different as it features both natural language descriptions (to support direct keyword search) and subjects based on bibliographic conventions. Indeed, the relevant subject terms selected by the indexer were first translated into the appropriate controlled vocabulary, and then organized into a compound subject (string), whose citation order follows the syntax of the indexing language.

The indexing language used for the project in question consists of: (1) a controlled vocabulary, (2) a syntax (citation order), and (3) sentences (indexing strings). In technical terms, this type of indexing is called *pre-coordination* as the relationships between concepts are established in advance, i.e. at the indexing stage. In pre-coordination, indexing offers a twofold advantage: (1) context, and (2) browsability. The latter can be implemented either as a hierarchy of strings (to meet the needs of expert users) or as a multidimensional space (to meet the needs of non-expert users). Pre-coordinated strings can be parsed according to the categories underlying the citation order for the sake of creating a multifaceted space of information that allows for post-coordination. Last but not least, pre-coordination is a helpful feature for improving proximity searches and ranking metrics [1].

As for Barthes' *punctum*, the analysis of the who, what, when and where present **in** an image is the responsibility of the beholder. Since in the era of Web 2.0 viewers are free to assign to visual content whatever tag in their opinion expresses the feelings, thoughts or memories arising in them when seeing an image, it is assumed that folksonomies and tag clouds are, among other things, capable of gathering about-ness statements as intended by Barthes.

3 The Concept of Facet

Facet analysis (FA) was originally conceived as a method for conceptual analysis of semantic elements based on a given set of principles and rules [8], [9] (pp. 299–326). From this standpoint, a *facet* simply represents a type of concept. Information scientists have (re)discovered facets only quite recently. Indeed, the large-scale application of FA to information architectures designed for the Web dates to early 2000s. From then onwards, facets have become the standard way in which e-commerce websites organize their content [28]. However, it must be said that the notion of facet underlying this ever-growing trend goes far beyond the original meaning and is indicative of a major conceptual shift wherein

a “generic term used to denote any component of a compound subject” [23] (p. 88) has in the end become any attribute of a resource. This widespread understanding of facets seems to reflect the way in which FA has been received in Northern America. Two cases in point are the *Faceted Access to Subject Terminology* (FAST) project by OPACs 2.0, featuring Endeca Guided Search [7], and the principles set out in the ANSI/NISO Z39.19-2005 (2010) standard, where one reads that: (1) the possible attributes or facets of a resource are subject (*indexing string*), author, location, form, language; and (2) the concept of facet basically corresponds to that of attribute in the computing field.

Unlike these examples, the notion of facet borrowed here follows that of the mathematician and librarian Ranganathan. Accordingly, indexing focuses on what an image depicts, while FA is the technique wherein terms or concepts of the same type (facets) are identified and grouped together. Therefore, the browse and search interface is the result of an analytical process in which facets (subject metadata) and the corresponding values are identified using natural language. Facets and values are then validated against the subject indexing tool devised for this project, which is also used as a conceptual grid to identify basic categories during the first steps of FA.

4 Information Exploration Architectures

There are two main ways of organizing information: hierarchical and faceted. A *hierarchical structure* consists of a classed system organized according to a top-down approach, starting from the most generic concept – the most top-level class – and then dividing it into sub-classes of increasing specificity. In general, a subject can be divided and subdivided according to different characteristics and levels of specificity. However, in practice, both division and specificity should reflect the distinctions and level of analysis by which the subject matter is, so to speak, naturally structured. For this reason, it is more convenient to develop a system on a pragmatic basis, rather than a philosophical one. The best approach is to compromise and seek the right balance between literary and user warrant. On the one hand, the structure should basically conform to the way in which the subject matter is represented in the relevant literature, but on the other hand it should be accessible in much the same way as users search for this type of information in the related literature and in online settings.

A good example briefly illustrating the pros and cons of this approach are those images whose subject falls under the category *Building(s)*. Its hierarchical arrangement is divided by place first, and then by building type and building parts, respectively. Put it another way, three characteristics of division are applied to this class in the following order: Place – B. type – B. parts. This, on the one hand, results in the inevitable scattering of the subordinate attributes (*distributed relatives*), but on the other hand it represents the prescribed and predictable sequence by which concepts are not only sorted but also accessed (*citation order*). In this respect, if the main concern is to first find buildings in a given area, and then their types and components, the citation order will perfectly

match the users' interest. On the contrary, if the initial focus is on types, in order to collect every piece of relevant information, users have to browse every single place. Therefore, when deciding on the order in which the attributes should be placed, one should carefully consider which order is more likely to fit the needs of the intended users, as well as which aspects are of primary importance and which ones can be distributed instead [9] (pp. 8–12). There are usually many aspects (facets) potentially relevant to users. Locking them into a rigid arrangement favoring one facet over the others results in a one-dimensional approach, and hence in a unidirectional search, that is very limiting.

Unlike the hierarchical approach, the faceted one relies on a bottom-up technique called facet analysis (FA). The process begins with grouping individual concepts into classes on the basis of some common characteristic. So far, nothing new in this. FA is not very different from the classic principle of logical division according to which each class is derived from the application of a single characteristic of division at a time. However, as already mentioned, some reasonable compromise or balance between literary and user warrant is necessary when choosing the characteristics for each class. Such characteristics are called facets and identify the properties used to model the information space of interest [29] (pp. 12–13). Facets are also referred to as conceptual dimensions of data or faceted metadata [33]. This approach seems to be a somewhat natural way of organizing information, since it offers the ability to accommodate on the same level all the facets (dimensions), or at least, the most relevant ones.

5 Search-and-Browse or Browse-and-Search Interface?

There are search situations in which non-expert users come with fuzzy information needs in mind. Since they lack background knowledge about the domain in question, the first problem to be tackled is how to help them get a better understanding of the domain and its conceptual organization. In practical terms, this means to find out what type of exploratory system is needed to achieve this goal.

It is assumed that the ultimate goal of exploratory search (ES) is to foster learning and investigation by combining analytical strategies (direct search) with browsing [18, 32]. Many applications make use of faceted search (FS) as a key technique for supporting ESs as FS combines these two components in a seamless fashion [28]. But how are these components combined together? And, what is more, how is the interaction between the two worked out? In a typical scenario, FS would begin with a keyword search: users have to formulate their information need as a query. Non-expert or naive users perform core exploratory tasks and issue short general queries returning large result sets [3]. This is when faceting is implemented to enhance the display of search results. Facets offer guidance to users in: (1) exploring the search space from different perspectives, and (2) narrowing down the result set based on relevant facets and associated values. This search-and-browse interaction pattern seems to be suitable for borderline lookup tasks [3].

Non-expert or casual users interact differently with cultural heritage information systems (CHISs) [20, 30, 31]. Coming without a specific information need

in mind, or with no need at all, these users just look around aimlessly, driven by the desire to discover something interesting, engaging or exciting. They interact with the information space in an exploratory fashion using browsing features first [30]. Therefore, to present a digital environment that is not only familiar and intuitive to a large majority of users, but also capable of providing handy orientation from the very beginning, is paramount.

A core assumption underlying browsing models is that metaphors based on physical spaces and resources are highly favored by both novice and expert users. These metaphors are likely to closely match users' mental models because they feel engaged in well-known physical environments such as the physical layout of a library [5, 24]. In fact, as shown in the Flamenco interface [33], faceted systems are multidimensional exploratory maps and, at the same time, spatial models where information is organized in much the same way as books are classified and placed on library shelves. Users get an overall view of the information space covered by the system, and then browse the content as if they were walking through the library shelves.

Faceted classification – taken in its strict meaning – represents a promising approach to the organization of information and construction of browse-and-search systems that may help newcomers in their initial interaction with digital cultural resources (cf. [8]). Moreover, faceted techniques can prompt the exploration tasks of casual users who approach digital cultural heritage content with the hedonic intent of finding something amusing, entertaining or enjoyable, and eventually discover that learning is fun too.

6 The Browse-and-Search Interface

The screenshot below (Fig. 2) shows the latest version of the user interface for an ongoing digital cultural heritage project called *Percorsi della memoria*, versione 2.0 [Paths of Memory, version 2.0] [11]. Its design largely follows the guidance and recommendations put forward for the *Flamenco Search Interface Project* [33], except for the faceted metadata. The center of the page is occupied by the set of results returned after selecting *Forme > Generi artistici* [Forms > Artistic genres]. The main faceted navigation system is located on the left side of the page. Its structure was built from the bottom up by first analyzing the subject terminology according to the grid of basic categories identified using the subject indexing tool devised for this project. The inner structure of each facet was subsequently organized using facet analysis. All facets are hierarchical, except for *Luoghi* [Places].

This system represents a knowledge structure that provides users with an overview of the information space based on its top-level semantic facets (the number of items contained in each facet is displayed next to the facet name). Clicking a facet displays the next level (checkbox/sub-facet pairs). In this case, the query preview shows not only the counts but also all the sub-facets nested under a facet, which are displayed after selecting the checkbox. Finally, the value level is reached by clicking the sub-facet (here the counts always refer to the

Forme > Generi artistici

The screenshot shows the following facets and results:

- Generi artistici (20)** [per categoria: cronologico]
 - Interni (18)**
 - Interni — Aquileia (1908)
 - Interni — Comons (1910)
 - Interni — Gorizia (1916)
 - Interni — Monte Santo (Slovenia) (1918)
 - Interni — Comons (1920-30)
 - Interni — Comons (1920-30)
 - Interni — Barbana (Grado) (1930-40)
 - Interni — Comons (1931)
 - Interni — Comons (1942)
- Filtri**
 - Categorie**
 - Interni (18)
 - Paesaggio (2)
 - Luoghi**
 - Aquileia (1)
 - Barbana (Grado) (1)
 - Brazzano (1)
 - Capriva del Friuli (3)
 - Comons (9)
 - Gorizia (3)
 - Monte Santo (Slovenia) (2)
 - Anni**
 - 1908 (1)
 - 1910 (1)
 - 1916 (1)
 - 1918 (1)
 - 1920-30 (2)

Fig. 2. Screenshot of the exploratory interface of *Percorsi della memoria*, version 2.0: set of results returned after selecting *Forme > Generi artistici* [Forms > Artistic genres]. Images are grouped by genre type and arranged in chronological order. On the right side of the page, there are facets that allow one to refine the set of results, and hence to contextualize the subject explored. The main faceted navigation system is located on the left side of the page.

number of matching results). When clicking a value, results are limited to those matching the corresponding value, as happens with a `WHERE` constraint, while selecting a checkbox/sub-facet pair allows the disjunctive selection of multiple values within a facet (ORed set).

The set of thumbnail images and captions displayed on the middle of the page is arranged in ascending chronological order to create a historical sequence of data. Many sets resulting from the selection of a sub-facet are instead classified according to the values present within each sub-facet. These categorized overviews are ordered alphabetically, while each sub-group is arranged chronologically. Since the size of the set is large, an additional faceting feature has been added: on the right side of the page, there are terms related to category type, time and space, which allow one to refine the results, and hence to contextualize the subject explored. This feature is always available to users so that they can filter sets exceeding 10 results.

Two additional key features are: (1) the user exploration path, which is shown at the top of the page; and (2) a separate line in the list below the main faceted navigation system displaying each exploration step (this list may serve as an exploration history as well). Last but not least, the interface remains the same throughout the user's journey. This browse-and-search framework model will be used for other similar collections of digital images such as the photographic collection of the Italian indologist Luigi Pio Tessitori [13].

7 Conclusion and Future Work

Non-expert or casual users browse digital cultural content prompted by a desire of finding something amusing, engaging or just interesting. In this regard, it is clear that better tools are needed to support this exploratory seeking behavior. Faceted classification seems to offer a good basis for developing browsing spaces where an initial journey for fun may become a journey for learning. The middle game phase of the information-seeking process tends to be the most problematic for users [33]. According to Jackson and colleagues [17], it is exactly in this phase that focused search plays a crucial role. However, further research is needed on how to insert direct keyword search in the middle game phase. It is expected that useful findings will emerge from a dedicated qualitative survey of user behavior which is still ongoing. Although the system described in this paper is still being developed, a working demo is available from <http://www.cataloguing-science.org/public/issrgo/ui.php>.

References

1. Pre- vs post-coordination and related issues (2007). https://www.loc.gov/catdir/cpsd/pre_vs_post.pdf
2. Akgül, G.B., Rubin, D.L., Napel, S., Beaulieu, C.F., Greenspan, H., Acar, B.: Content-based image retrieval in radiology: current status and future directions. *J. Digit. Imaging* **24**(2), 208–222 (2011)
3. Athukorala, K., Glowacka, D., Jacucci, G., Oulasvirta, A., Vreeken, J.: Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks. *J. Assoc. Inf. Sci. Technol.* **67**(11), 2635–2651 (2016)
4. Barthes, R.: *La chambre claire. Note sur la photographie*. Gallimard-Seuil (1980)
5. Bates, M.J.: The design of browsing and berrypicking techniques for the online search interface. *Online Rev.* **13**(5), 407–424 (1989)
6. Bérengère, C.: La boîte noire de Daguerre: un nouvel espace-temps. *Image Narrat.* **23** (2008). <http://www.imageandnarrative.be/inarchive/Timeandphotography/chapuis.html>
7. Breeding, M.: Chapter 3: Endeca. *Libr. Technol. Rep.* **4** (2007). <https://journals.ala.org/index.php/ltr/article/view/4540/5333>
8. Broughton, V.: The need for a faceted classification as the basis of all methods of information retrieval. *ASLIB Proc.* **58**(1–2), 49–72 (2006)
9. Broughton, V.: *Essential Classification*. Facet Publishing, London (2015)
10. Christensen, H.D.: Rethinking image indexing. *J. Assoc. Inf. Sci. Technol.* **68**(7), 1782–1785 (2017)
11. Cuna, A.: Immagini e web: percorsi della memoria 2.0. *Tafer J.* **61** (2013). <http://www.taferjournal.it/2013/07/03>
12. Dean, J., Barroso, L.A.: The tail at scale. *Commun. ACM* **56**(2), 74–80 (2013)
13. Freschi, F.: Luigi Pio Tessitori’s unpublished works in the archivio Peano. In: Tessitori and Rajasthan, pp. 197–219. *Soc. Ind. “L.P. Tessitori”* (1999)
14. Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T.: The vocabulary problem in human-system communication. *Commun. ACM* **30**(11), 964–971 (1987)
15. Harpring, P.: The language of images: enhancing access to images by applying metadata schemas and structured vocabularies. In: *Introduction to Art Image Access*, pp. 20–39. Getty Research Institute (2002)

16. Hearst, M.A.: Search User Interfaces. Cambridge University Press, Cambridge (2009)
17. Jackson, A., Lin, J., Milligan, I., Ruest, N.: Desiderata for exploratory search to web archives in support of scholarly activities. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, vol. 16, pp. 103–106 (2016). https://cs.uwaterloo.ca/~jimmylin/publications/Jackson_etal_JCDL2016.pdf
18. Marchionini, G.: Exploratory search: from finding to understanding. *Commun. ACM* **49**(1), 41–46 (2006)
19. Markey, K.: Computer-assisted construction of a thematic catalog of primary and secondary subject matter. *Vis. Resour.* **3**, 16–49 (1983)
20. Mayr, E., Federico, P., Miksch, S., Schreder, G., Smuc, M., Windhager, F.: Visualization of cultural heritage data for casual users. <http://publik.tuwien.ac.at/files/PubDat.250950.pdf>
21. Mitchell, W.J.T.: *Picture Theory*. University of Chicago Press, Chicago (1995)
22. Panofsky, E.: *Studies in Iconology: Humanistic Themes in the Art of the Renaissance*. Oxford University Press, Oxford (1939)
23. Ranganathan, S.R.: *Prolegomena to Library Science*. Asia Publishing, New York (1967)
24. Rimmer, J., Warwick, C., Blandford, A., Gow, J., Buchanan, G.: An examination of the physical and the digital qualities of humanities research. *Inf. Process. Manag.* **44**(3), 1374–1392 (2008)
25. Shatford, S.: Analyzing the subject of a picture: a theoretical approach. *Cat. Classif. Q.* **6**(3), 39–62 (1986)
26. Svenonius, E.: *The Intellectual Foundation of Information Organization*. MIT Press, Cambridge (2000)
27. Tsai, C.: A review of image retrieval methods for digital cultural heritage resources. *Online Inf. Rev.* **31**(2), 185–198 (2007)
28. Tunkelang, D.: *Faceted Search*. Morgan & Claypool, San Rafael (2009)
29. Vickery, B.C.: *Faceted Classification: A Guide to Construction and Use of Special Schemes*. Aslib, London (1960)
30. Villa, R., Clough, P., Hall, M., Rutter, S.: Search or browse? Casual information access to a cultural heritage collection. In: *EuroHCIR 2013* (2013)
31. Gäde, M., Hall, M., Huurdeman, H., Kamps, J., Koolen, M., Skov, M., Toms, E., Walsh, D.: Supporting complex search tasks. In: Hanbury, A., Kazai, G., Rauber, A., Fuhr, N. (eds.) *ECIR 2015*. LNCS, vol. 9022, pp. 841–844. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16354-3_99
32. White, R.W., Roth, R.A.: *Exploratory Search: Beyond the Query-Response Paradigm*. Morgan & Claypool, San Francisco (2009)
33. Yee, K.P., Swearingen, K., Li, K., Hearst, M.: Faceted metadata for image search and browsing. In: *Proceedings of the SIGCHI Conference on the Human Factors in Computing Systems* (2003)

Library Data Integration: The CoBiS Linked Open Data Project and Portal

Luisa Schiavone¹(✉), Federico Morando², Davide Allavena²,
and Giorgio Bevilacqua³

¹ INAF Turin Astrophysical Observatory, Pino Torinese, Italy

luisa.schiavone@inaf.it

² Synapta Srl, Turin, Italy

info@synapta.it

³ Department of Humanities, University of Turin, Turin, Italy

338853@edu.unito.it

Abstract. The CoBiS is a network formed by 65 libraries. The project is a pilot for Piedmont aiming to provide the libraries with an infrastructure for LOD publishing, creating a triplification pipeline designed to be easy to automate and replicate. This was realized with open source technologies, such as the TARQL and JARQL tools that use SPARQL queries to describe the conversion of tables (CSV) or trees (JSON) into graphs (RDF data). The first challenge consisted in making possible the dialog of heterogeneous data sources, coming from four different library applications and different types of data. As a second step, the information contained in the catalogs was interlinked with external data sources.

1 Introduction

The **CoBiS** (Coordinamento delle Biblioteche Speciali e Specialistiche di Torino i.e. *Coordination of Special and Specialized Libraries of Turin*) is an informal network of 65 libraries, collaborating to provide continuing professional development and to offer a better service to their users. CoBiS libraries are heterogeneous from many points of view: holdings, cataloguing software and OPACs. The **LOD project** started in 2015 as a training program, in collaboration with Prof. Vivarelli from the University of Turin. The program was divided in various topics: copyright, collaboration between libraries and Wikipedia, (Linked) Open Data.

2 Purpose

Six libraries from the CoBiS decided to participate in a **pilot project** with the purpose of **providing a unique access point to the collections of CoBiS libraries**. CoBiS bibliographic data were divergent from different perspectives: file formats, semantic frameworks and authority files.

Linked Open Data technologies provide the means to engage such interoperability problems, both from a technical and a semantic point of view. Many national libraries (e.g. BNE [1] and BNF [2]) have converted their catalogues from MARC to RDF and have published their data also through a SPARQL endpoint.

In Italy, the SHARE Catalogue project¹ has converted MARC records from multiple university libraries in RDF; however, no SPARQL endpoint is available to query the database.

By developing and implementing effective tools and procedures, we were able to convert CoBiS datasets, composed of different data formats, into RDF, to interlink them with external authoritative sources (Wikidata, VIAF and InternetArchive) and to publish them as Linked Data, giving also access to the enriched dataset via a SPARQL endpoint.

This work led CoBiS datasets to become a connecting piece in the Linked Open Data Cloud; as a result, data are not only becoming more interoperable among them, but they are also open in order to facilitate the collaboration with online communities.

3 The Project

With respect to the **Linked Open Data stack**, Fig. 1 shows an overall picture of the **project**.

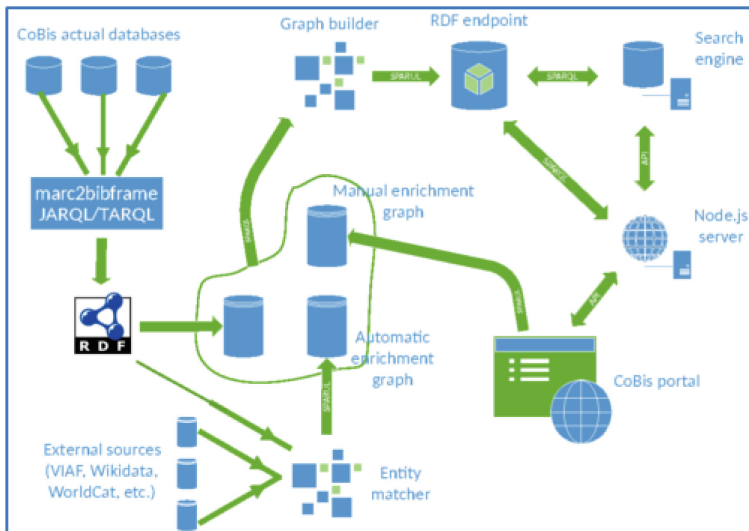


Fig. 1. The project architecture

¹ <http://catalogo.share-cat.unina.it/sharecat/clusters>.

The first activity of the project consisted in transforming bibliographic data into RDF triples.

Created using various cataloguing softwares, CoBiS data are encoded in different file formats (CSV, XML and JSON) and are structured according to different data models (MARC 21, UNIMARC, Dublin Core OAI-PMH).

In order to define a unique data model we used two main **ontologies**: Schema.org and Bibframe, the latter being developed by the Library of Congress with the specific aim to facilitate the necessary transition from the traditional catalographic system (based on MARC) to a bibliographic *environment* integrated in the *web of data* [3]. We also used selected properties from RDFS, OWL, DCTerms, FOAF, and Culturalis², with the purpose of providing more semantic interoperability.

During the first phase of the project, we tried to convert data in RDF triples using the Library of Congress tool `marc2bibframe`³ to process MARC 21 and using the RML mapping language for all the other formats.

RML is a *generic mapping language defined to express customized mapping rules from heterogeneous data structures and serializations to the rdf data model* [4]. The RML conceptual model, based on R2RML W3C standard⁴, perfectly fitted with the needs of the project, but had some limitations with respect to technical performances, i.e. processing time, and the verbosity of the mapping language.

When tabular data were available, we experimented the TARQL tool, described by its developers as *SPARQL for Tables: a command-line tool for converting CSV files to RDF using SPARQL 1.1 syntax*⁵.

TARQL proved to be both technically efficient and not very time-consuming in terms of writing new mappings, thanks to the use of the standard SPARQL 1.1 syntax, with all its features.

Since many of our input data had a tree-like structure (XML or JSON format) Synapta, in a joined effort with other open source developers from FactsMission⁶, realized and published JARQL, a new open source software tool for converting JSON data to RDF. As TARQL, from which it takes inspiration, JARQL uses the SPARQL 1.1 syntax and constructs queries to describe its mappings⁷.

By improving those technical tools, we were able to define a **triplification pipeline** where data are extracted from local sources, mapped (using a SPARQL query) to selected ontologies, and converted into an RDF graph, which is periodically updated. A sample JSON input and the related SPARQL mapping is shown in Fig. 2.

² <http://culturalis.org/>.

³ <https://github.com/lcnetdev/marc2bibframe>.

⁴ <https://www.w3.org/TR/r2rml/>.

⁵ <http://tarql.github.io/>.

⁶ <https://factsmission.com/>.

⁷ <http://jarql.linked.solutions/>.



Fig. 2. JSON input and related JARQL mapping using SPARQL 1.1 syntax

According to the W3C recommendations [5], CoBiS entities (books and authors) are unambiguously identified by URIs, so that they can be connected to external sources.

CoBiS libraries were using different authority files (e.g. SBN or local systems), but the graph structure of RDF supported the generation of a unified Authority file. Interlinks to external sources, like VIAF and Wikidata, helped us to identify and de-duplicate authors under a shared procedure.

For **link generation** [6] we used both automatic algorithms and manual approaches. In order to minimize false positives, automatic matches proceeded mostly through SBN identifiers: in this way, the newly created graph was inter-linked with VIAF and Wikidata.

Due to the size of the VIAF database and to the absence of a public SPARQL end-point, we had to recreate locally a Linked Data graph from the VIAF RDF dump.

In order to support manual matches, we exploited **OLAF (Open Linked Authority File)**⁸, our crowd-sourcing interface for creating an authority file based on SPARQL queries.

By analyzing different RDF-structured databases, OLAF suggests potential relations of identity between similar entities (see Fig. 3).

Candidate matches are submitted to domain-experts and, if validated, they are annotated in the CoBiS RDF graph, using the owl:sameAs property. The interaction between automatic computation and manual validation improves the quality and the **reliability** [7] of the data, by allowing domain-experts to directly contribute to the Linked Open Data Cloud.

⁸ <https://olaf.synapta.io/>.

The screenshot shows the 'Olaf Linked Authority File - CoBiS' interface. At the top, there is a blue header with the text 'Olaf Linked Authority File - CoBiS'. Below this, the main content is organized into several sections:

- Galilei Person:** A green box containing the name 'Galilei', the title 'Person', and a biographical summary: '1564-1642 // Matematico, astronomo e filosofo, nato a Pisa e morto ad Arcetri (FI). Fu professore a Pisa e a Padova, matematico e filosofo del Granuova di Toscana (1510), accademico dei Lincei.'
- Antologia delle opere maggiori:** A green box listing 'Scritti vari', 'Studi sulla Divina Commedia', 'Il saggiaiore', and 'Una lettera inedita di Galileo Galilei'.
- Buttons:** A grey 'SALTA' button and an orange 'NESSUNO DI QUESTI' button.
- Vincenzo Galilei:** A white box with a green footer. It lists 'compositore e teorico musicale italiano', birth 'Nascita: 1520-4-3', and death 'Morte: 1591-7-2'. It includes links for 'WIKIDATA', 'WIKIPEDIA IN ITALIANO', and 'WIKIPEDIA IN INGLESE'. The footer says 'SONO IO!'.
- Galileo Galilei:** A white box with a green footer. It features a portrait of Galileo Galilei, lists 'scienziato italiano', birth 'Nascita: 1564-2-25', and death 'Morte: 1642-1-8'. It includes links for 'WIKIDATA', 'WIKIPEDIA IN ITALIANO', and 'WIKIPEDIA IN INGLESE'. The footer says 'SONO IO!'.
- Michelangelo Galilei:** A white box with a green footer. It lists 'compositore italiano', birth 'Nascita: 1575-12-18', and death 'Morte: 1631-1-3'.
- Galilei:** A white box with a green footer. It lists 'pagina di disambiguazione di un progetto Wikimedia' and includes links for 'WIKIDATA' and 'WIKIPEDIA IN ITALIANO'.

Fig. 3. The *Galilei* search in OLAF

Within the overall interlinking process, **Wikidata** is assuming the role of meta-data hub: because of its constant growing and the dynamism of the *wiki* approach [8], its entities are strongly interlinked with VIAF and other databases (e.g. national bibliographic and biographic dictionaries, disciplinary databases, etc.).

On a scenario where the cataloging process is evolving into aggregating shared and interlinked data [9], librarians (especially those of small institutions not contributing to VIAF) are encouraged to use Wikidata as authority file [10], both by connecting existing data and creating new items about authors not yet included in Wikidata (using its publicly editable graphic interface). In those cases, such a practice would minimize redundancy and the overall cost of authority record creation, thus increasing the efficiency of bibliographic record production and maintenance [11].

4 The Portal

The CoBiS LOD Project portal (<http://dati.cobis.to.it>) is online with its full Linked Data stack, including a public SPARQL end-point configured to support federated queries (<http://dati.cobis.to.it/sparql>) a full dump of the RDF data, etc.

In the author's page, dynamically generated through SPARQL queries, you have biographical information and a list of interlinked resources coming from Wikidata and other bibliographic repositories (VIAF, Wikidata, LoC, Deutsche National Bibliothek GND, Bibliothèque Nationale de France BNF, Servizio Bibliotecario Nazionale SBN, Dizionario Biografico degli Italiani DBI).

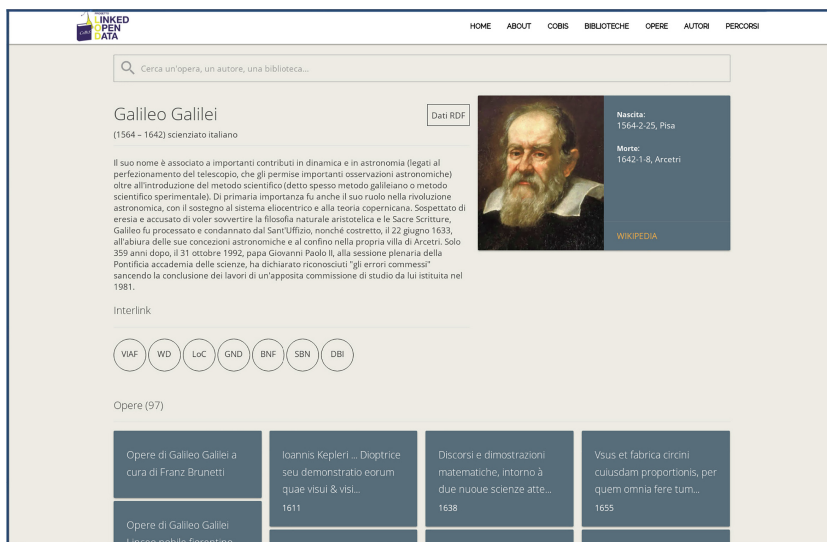


Fig. 4. The *Galileo Galilei* page

On the right there is an infobox with authors data and an image, both fetched live from Wikidata leveraging Linked Data. Clicking on the RDF button, all the triples of the resource can be directly viewed.

At the bottom of the page, all the author's **books inside the CoBiS database** are shown (see Fig. 4). To explore information on such books, you can click one of the boxes or you can use the search bar, looking for a title which is not listed⁹.

Figure 5 shows an **example search for the *Dialogo***. On the left side of the page, you see bibliographic details with a collection of interlinked resources. Exploiting the power of Open Data, we are also able to read the **Internet Archive digital copy of the book** (Fig. 6). A physical copy of the book is

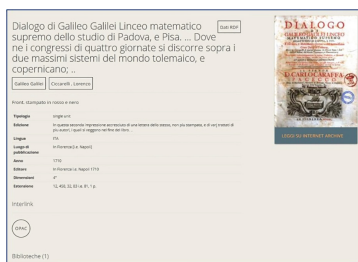


Fig. 5. The Galilei's *Dialogo* file in the CoBiS database

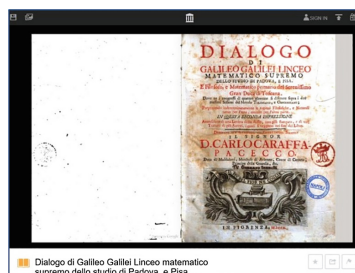


Fig. 6. The Galilei's *Dialogo* in the internet archive

⁹ The search bar is powered by Solr - <https://lucene.apache.org/solr/>.

available in some CoBiS libraries. All details coming from the individual OPACs can be shown by clicking the **OPAC button** (Fig. 5).

5 Towards New Challenges

The pilot is only an initial step.

CoBiS plan to include more libraries in a second phase of the project, converting their data into Linked Open Data and linking them to the Linked Open Data Cloud and to other online resources, such as Internet Archive and Wikipedia.

Thanks to linked data, it is possible to query for common items in the CoBiS catalogue and in other major National Libraries catalogues exposing a SPARQL endpoint.

Performing the query in Fig. 7 on the CoBiS SPARQL endpoint¹⁰ towards the French National Library SPARQL endpoint¹¹ returns a subset of shared books.

```
PREFIX bnf-onto: <http://data.bnf.fr/ontology/bnf-onto/>
PREFIX schemaorg: <http://schema.org/>

select ?cobisInstance ?BNFInstance
where {
  ?cobisInstace <http://schema.org/isbn> ?isbnCobis .
  service <http://data.bnf.fr/sparql> {
    {
      select ?isbn ?BNFInstance
      where {
        ?BNFInstance <http://data.bnf.fr/ontology/bnf-onto/isbn> ?isbn .
      } LIMIT 10000
    }
  }
  FILTER(REPLACE(STR(?isbn), "-", "" ) = ?isbnCobis)
} LIMIT 100
```

Fig. 7. The performed query

We also aim to improve the interlinking, so as to link CoBiS data to other online open data, with specific regard to Wikidata and VIAF and to the most important international projects, such as the linked open data portals of the French and Spanish National Libraries and of the Library of Congress in Washington DC.

¹⁰ <http://dati.cobis.to.it/sparql>.

¹¹ <http://data.bnf.fr/sparql>.

6 Acknowledgements

This pilot project would not have taken place without the aid of the following organizations, that we want to thank:

- *Regione Piemonte* for having believed in the project and for the contribution provided;
- *Politecnico di Torino, Management Committee of the Fund for Development of Research and Education in Information and Communication Technologies* for the financial support to the initial phase of the project;
- *Politecnico di Torino, Nexa Center for Internet and Society (DAUIN)* for the collaboration;
- *Synapta* for the technical realization;
- all the participating Institutes: National Institute for Astrophysics (INAF), Turin Academy of Sciences, Olivetti Historical Archives Association, Alpine Club National Library, Deputazione Subalpina di Storia Patria, National Institute for Metrological Research (INRIM).

References

1. Vila-Suero, D., Villazón-Terrazas, B., Gómez-Pérez, A.: datos.bne.es: A library linked dataset. *Semant. Web* 4(3), 307–313 (2013)
2. Simon, A., Wenz, R., Michel, V., Di Mascio, A.: Publishing bibliographic records on the web of data: opportunities for the BnF (French National Library). In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) *ESWC 2013. LNCS*, vol. 7882, pp. 563–577. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38288-8_38
3. Miller, E., et al.: Bibliographic framework as a web of data: linked data model and supporting services. Library of Congress, Washington DC (2012)
4. Dimou, A., Sande, M.V., Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R.: RML: A generic language for integrated RDF mappings of heterogeneous data. In: Bizer, C., Heath, T., Auer, S., Berners-Lee, T. (eds.) *Proceedings of the 7th Workshop on Linked Data on the Web* (2014)
5. *LinkedData*. <https://www.w3.org/DesignIssues/LinkedData.html>
6. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. *IJISWIS* 5(3), 1–22 (2009)
7. Guerrini, M., Possemato, T.: *Linked data per biblioteche, archivi e musei*. Editrice Bibliografica, Milano (2015)
8. Martinelli, L.: Wikidata: la soluzione wikimediana ai linked open data. *AIB Studi* 56(1), 75–85 (2016)
9. Bianchini, C.: RDA e la sfida del web semantico. In: De Castro, F. (ed.) *Il punto sul Servizio Bibliotecario Nazionale e le sue realizzazioni nel Friuli Venezia Giulia*, pp. 197–206. EUT Edizioni Università di Trieste, Trieste (2014)
10. Guerrini, M.: La filosofia open: paradigma del servizio contemporaneo. *Biblioteche oggi* 35, 12–21 (2017)
11. Library of Congress Working Group on the Future of Bibliographic Control: *On the Record: Report of the Library of Congress Working Group on the Future of Bibliographic Control*. Library of Congress, Washington DC (2008)

A Software Architecture for Narratives

Carlo Meghini, Valentina Bartalesi^(✉), Daniele Metilli, and Filippo Benedetti

Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" – CNR,
Pisa, Italy

{carlo.meghini, valentina.bartalesi, daniele.metilli,
filippo.benedetti}@isti.cnr.it

Abstract. The current Digital Libraries (DLs) usually return as answer of a user's query a ranked list of the resources included in the DLs but no semantic relation among the resources are reported. Using the Semantic Web technologies it is possible to improve these search functionalities introducing *narratives* as new search method. As *narratives* we intend semantic networks of events that are linked to the objects of the DLs and are endowed with a set of semantic relations that connect an event to another. These semantic networks may help the users to obtain a more complete knowledge on the subject of their searches. In this paper, we present a software architecture for building narratives in order to introduce them in DLs. Our architecture is composed of several tools (automatic and semi-automatic tools) for creating, storing and visualizing narratives. When possible, we reused open source components already available on-line, and for the software we developed, we freely distribute it for research aims.

Keywords: Software architecture · Semantic Web
Semantic reasoner · Narratives · OWL · Digital Libraries

1 Introduction

The search functionalities of the current Digital Libraries (DLs) are usually basic systems that answer to a user's query, expressed in natural language, with a ranked list of the resources included in the DLs but no semantic relations among the returned objects are reported. The Semantic Web [3], and the Linked Open Data [11] paradigm, can overcome the limitations of these search functionalities. The long-term aim of our research is to develop and integrate in DLs a new search method, using the semantic Web technologies: the *narrative*. We intend narratives as semantic networks composed of events that are linked to the objects of the DL and are endowed with a set of semantic relations connecting these events, i.e. actions or occurrences taking place at a certain time at a specific location. In our vision, instead of list of objects, DLs should provide the narratives as answers of the queries, which could be useful for users in order to obtain a more complete knowledge on the subject of their searches. To reach this aim, we developed a software architecture that allows to create narratives

using the Semantic Web technologies. This architecture is composed of a set of tools (automatic and semi-automatic tools) for creating, storing, querying, and visualizing narratives. The stored knowledge is formally represented following an OWL ontology for representing narratives [1] we developed, encoded in the OWL 2 DL profile [5]. In order to maximize its interoperability, our ontology was developed as an extension of the CIDOC CRM standard ontology [4].

We have created some narratives using this architecture. In particular, two biographical narratives were produced by a Digital Humanities researcher at the Italian National Research Council (CNR): (i) on the life of Dante Alighieri¹, the major Italian poet of the Middle Age; (ii) on the life of the Austrian painter Gustav Klimt². The third narrative was developed by a researcher in Computational Biology at the CNR to narrate the discoveries related to the giant squid³.

Several components of our architecture are already developed and open source, thus we reused them. For what regards the software we developed, we distribute it freely for research aims.

2 Architecture

This section describes our current architecture for the representation of narratives. Figure 1 shows the architecture, whose main components are the following:

1. *a narrative-building tool*. It is used for creating, modifying or visualizing a narrative, possibly representing knowledge that has been derived by reading some texts. The user operates through the Graphical User Interface (GUI) of the narrative-building tool, by manually inserting the narrative data and, at the same time, importing resources from Wikidata⁴. The created narrative is stored as an intermediate JSON representation⁵;
2. *an OWL triplifier*. Once the narrative is complete, the corresponding JSON representation is given as input to the Java Triplifier. The triplifier transforms the JSON file into an OWL (Web Ontology Language) ontology encoded as an RDF graph, using the OWL API library [6]. The organization of the knowledge in the graph follows the structure defined in the ontology for narratives we developed [1];
3. *a semantic reasoner*. It is used by the triplifier to infer new knowledge. The triplifier takes as input also a file with SWRL rules [7] that are used by the reasoner to support the temporal reasoning on the narrative;
4. *a triple store*. The triplifier stores the resulting graph, expanded with inferences produced by both the reasoner and the SWRL rules, into a Blazegraph triple store⁶;

¹ <https://dlnarratives.eu/timeline/dante.html>.

² <https://dlnarratives.eu/timeline/klimt.html>.

³ <https://dlnarratives.eu/timeline/squid.html>.

⁴ <https://www.wikidata.org>.

⁵ <http://json.org/>.

⁶ <https://www.blazegraph.com/>.

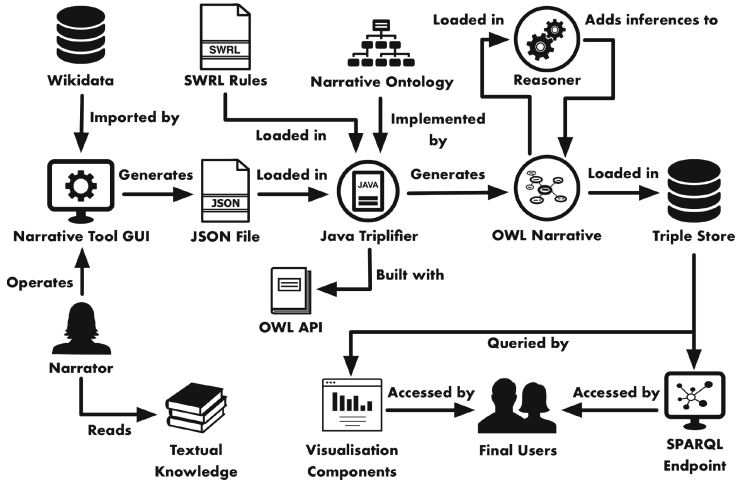


Fig. 1. The schema of our architecture.

5. *a visualization interface*. Finally, the user can access the knowledge stored in the triple store through a Web interface. The knowledge is extracted using SPARQL queries [10] and shown using graphic libraries. A review of approaches relevant to our study is reported in [8].

2.1 Narrative-Building Tool

In order to facilitate the creation of a narrative and its semantic representation by the narrator, we built a web-based *narrative-building tool*. The tool is built with HTML5⁷ and JavaScript (ECMAScript6⁸), using the jQuery⁹, jQuery UI¹⁰, Bootstrap¹¹, and Typeahead.js¹² libraries.

The main interface of the tool is based on simple drag-and-drop metaphors, allowing the user to create events and drag the appropriate entities that compose them from a list of entities (e.g. location, person, object) automatically extracted from the Wikidata knowledge base through its SPARQL endpoint. The entities are color-coded according to their class, i.e. person, organization, place, object, concept, or work. These are linked to the corresponding CRM classes by using a specific mapping we defined. The interface also allows the user to link together events by using two different semantic relations: (i) the mereological (part-of)

⁷ <https://www.w3.org/TR/html5/>.

⁸ <http://www.ecma-international.org/ecma-262/6.0/>.

⁹ <https://jquery.com>.

¹⁰ <https://jqueryui.com>.

¹¹ <https://getbootstrap.com>.

¹² <https://typeahead.js.org>.

Fig. 2. The interface of the NBVT.

relation, or (ii) the causal dependency relation. A view of the NBVT interface is reported in Fig. 2.

The tool is available online¹³, it is open-source and released under the GPLv3 license¹⁴. As the user inserts knowledge in the interface of the tool, the data is stored into a CouchDB¹⁵ database, using the PouchDB¹⁶ library to interface with it. This allows automatic saving of the data, revisioning, and synchronization between a local and remote database. Finally, the resulting timeline of events is exported in the JSON format.

2.2 OWL Triplifier

The knowledge exported from the tool in JSON format is subsequently imported into a Java-based triplifier. The triplifier makes use of the OWL API library to define an ontology model. Then, it loads the JSON file and converts it to an intermediate Java representation. Then the OWL API library takes as input this representation that is used for populating the model. The triplifier also imports some SWRL rules. The Semantic Web Rule Language (SWRL) is a proposed language for the Semantic Web that can be used to express rules as well as logic, combining OWL DL or OWL Lite with a subset of the Rule Markup Language. We added SWRL rules to overcome the limitations of the OWL 2 DL

¹³ <https://dlnarratives.eu/tool.html>.

¹⁴ <https://www.gnu.org/licenses/gpl-3.0.en.html>.

¹⁵ <http://couchdb.apache.org>.

¹⁶ <https://pouchdb.com>.

about: (i) the definition of a relation as simultaneously transitive and irreflexive, as in the case of the part-of and causality relations, (ii) the definition of a relation as simultaneously transitive and disjoint [9], as in the case of the temporal relation, (iii) the implication between part-of and temporal relations and between causality and temporal relations. For this reason, we use SWRL rules and OWL 2 DL axioms simultaneously. The SWRL rules were produced using a software developed by Batsakis et al. [2] for what concerns the temporal relations. The implications between part-of and temporal relations and between causality and temporal relations were defined by writing the SWRL rules manually. The SWRL rules are taken as input by the triplifier as an OWL file. The result of the process of triplifier is an OWL graph that represents the narrative, exportable in RDF/XML¹⁷, Turtle¹⁸, or several other syntaxes¹⁹.

2.3 Reasoner

At this point, a reasoning is performed on the knowledge in order to perform consistency checks and make inferences. The reasoner we adopted is Openllet²⁰ version 2.6. The main reasons for this choice are the following: (i) Openllet supports all the features of OWL 2 DL; (ii) it fully supports SWRL rules; (iii) it is Java-based and easily integrated with OWL API²¹; (iv) it is an open source software actively maintained. Openllet provides functionality to check consistency of ontologies, compute the classification hierarchy, explain inferences, the graph is stored into a triple store.

2.4 Triple Store

The knowledge is exported to a triple store. The triple store we chose is Blaze-graph²². Blazegraph is a standards-based, high-performance, scalable, open-source graph database. Written entirely in Java, the platform supports the RDF data model and the SPARQL 1.1 family of specifications, including Query, Update, Basic Federated Query, and Service Description. The knowledge stored in Blazegraph is shown to the user through a visualization interface. We implemented SPARQL queries to retrieve this knowledge from the triple store.

2.5 Visualization Interface

First of all, in order to give a complete overview of the narrative, the events were placed on a timeline. We used TimelineJS library²³ for the implementation. For

¹⁷ <https://jena.apache.org/documentation/io/rdf-output.html>.

¹⁸ <https://www.w3.org/TR/turtle/>.

¹⁹ <https://jena.apache.org/documentation/io/rdf-output.html>.

²⁰ <https://github.com/Galigator/openllet>.

²¹ <https://owls.github.io/owlapi/>.

²² <https://www.blazegraph.com/>.

²³ <https://timeline.knightlab.com/>.

each event on the timeline, the more meaningful information is reported, i.e. title, date, primary sources, related digital objects, related images. Events occurred at the same time are allowed and visualised on a timeline. Figure 3 shows an event of the timeline of Dante Alighieri's life.

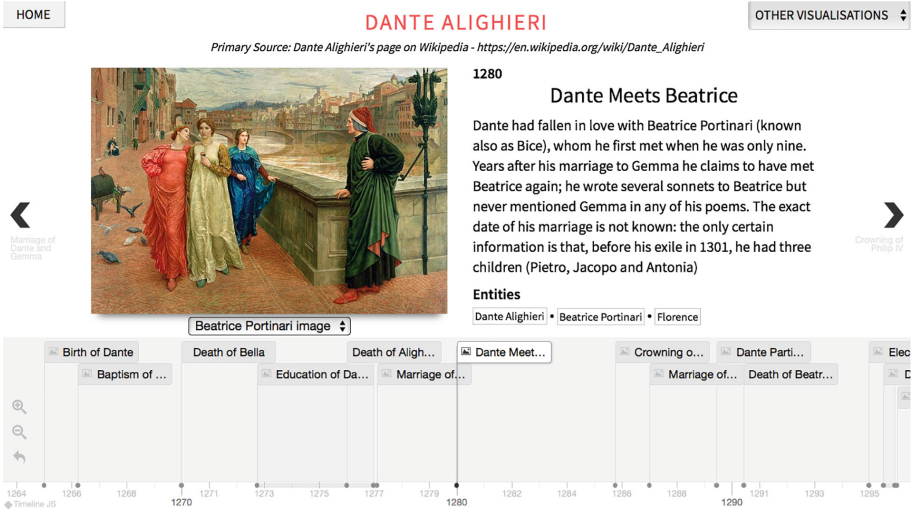


Fig. 3. An event of Dante Alighieri's life on the timeline.

Another requirement for the tool is the visualization of the entities that compose each event. To this aim, a SPARQL query to get this information from the knowledge base was implemented. This query retrieves, for each event title, the names and IRIs of the corresponding entities. The `vis.js`²⁴ JavaScript library was used to implement the visualization. One of the most important requirements for a scholar who studies historical events, is the knowledge of their primary sources. For each event of the narrative, the tool allows to visualize the primary sources and in particular the title and the author of a primary source, the textual fragment of the primary source that describes the event, the reference of the textual fragment. This information is visualized in tabular format. Finally, the user has the possibility to visualize all events that occurred in a specified period of time. Upon specifying the desired period, the user can freely insert the dates using a widget to select a full date or the year only. The results of the query are shown in form of table, where for each event its dates are shown.

It is possible to explore the visualization interface on-line²⁵, browsing the three narratives that are available on our Web site²⁶.

²⁴ <http://visjs.org>.

²⁵ <https://dlnarratives.eu/narratives.html>.

²⁶ <https://dlnarratives.eu>.

3 Conclusions and Future Work

In this paper we have presented a software architecture for building narratives using the Semantic Web technologies. In this context, we intend narratives as semantic networks of events linked to each other and to digital objects by semantic relations. In order to represent the knowledge we have developed an OWL ontology for narratives as an extension of the CIDOC CRM standard vocabulary.

Where possible, to develop this architecture, we have reused software open source already available. For what regards the software we developed, it is freely distributed, released under the GPLv3 license.

The long-term goal of our study is introducing the narrative as new first-class search functionality of digital libraries. As output of a query, this new search functionality should not only return a list of objects but it should also present one or more narratives on the topic of the search. The architecture presented in this paper would be used to create narratives that later could be imported and shown in the DL interfaces.

References

1. Bartalesi, V., Meghini, C., Metilli, D.: Steps towards a formal ontology of narratives based on narratology. In: OASIS-OpenAccess Series in Informatics, vol. 53. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2016)
2. Batsakis, S., Petrakis, E., Tachmazidis, I., Antoniou, G.: Temporal representation and reasoning in OWL 2. *Semant. Web* **8**(6), 981–1000 (2016)
3. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic Web. *Sci. Am.* **284**(5), 28–37 (2001)
4. Doerr, M.: The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI Mag.* **24**(3), 75 (2003)
5. Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F., Rudolph, S.: OWL 2 Web ontology language primer. *W3C Recomm.* **27**(1), 123 (2009)
6. Horridge, M., Bechhofer, S.: The OWL API: a Java API for working with OWL 2 ontologies. In: Proceedings of the 6th International Conference on OWL: Experiences and Directions, Aachen, Germany, OWLED 2009, vol. 529, pp. 49–58. CEUR-WS.org (2009). <http://dl.acm.org/citation.cfm?id=2890046.2890052>
7. Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Groszof, B., Dean, M., et al.: SWRL: a semantic Web rule language combining OWL and RuleML (2004)
8. Meghini, C., Bartalesi, V., Metilli, D.: Using formal narratives in digital libraries. In: Grana, C., Baraldi, L. (eds.) IRCDL 2017. CCIS, vol. 733, pp. 83–94. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68130-6_7
9. Motik, B., Patel-Schneider, P.F., Parsia, B., Bock, C., Fokoue, A., Haase, P., Hoekstra, R., Horrocks, I., Ruttenberg, A., Sattler, U., et al.: OWL 2 Web ontology language: structural specification and functional-style syntax. *W3C Recomm.* **27**(65), 159 (2009)
10. Prud'hommeaux, E., Seaborne, A.: SPARQL Query Language for RDF. *W3C Recomm.* (2008). <https://www.w3.org/TR/rdf-sparql-query/>
11. Yu, L.: Linked open data. In: Yu, L. (ed.) A Developers Guide to the Semantic Web, pp. 409–466. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-15970-1_11

Thirty Years of Digital Libraries Research at the University of Padua: The Systems Side

Maristella Agosti, Giorgio Maria Di Nunzio, Nicola Ferro,
and Gianmaria Silvello^(✉)

University of Padua, Padua, Italy
{maristella.agosti,giorgiomaria.dinunzio,nicola.ferro,
gianmaria.silvello}@unipd.it

Abstract. For the thirty years of the Information Management Systems (IMS) research group of the University of Padua, we report the main and more recent contributions of the group to the field of Digital Library Systems. In particular, we briefly describe the systems designed and developed by members of the group in the context of research infrastructures, digital archives, digital linguistics and scientific data.

1 Introduction

Digital libraries have contributed to supporting the creation of innovative applications and services to access, share and search our cultural heritage. One of the most important contributions of digital libraries is to make available collections of digital resources from different cultural institutions such as *libraries*, *archives* and *museums*, to make them accessible in different languages and to provide advanced services over them. Digital libraries are heterogeneous systems with functionalities that range from data representation to data exchange and data management. Furthermore, digital libraries are meaningful parts of a global information network which includes scientific repositories, curated databases and commercial providers. All these aspects need to be taken into account and balanced to support final users with effective and interoperable information systems.

In the last thirty years the Information Management Systems (IMS) research group of the University of Padua has contributed to the design and development of diverse digital library systems contributing to the foundations of the field by providing an interoperability layer between the DELOS model and the 5S model (Sect. 2), to research infrastructures with the CULTURA environment (Sect. 3), to digital archives with the SIAR system (Sect. 4), to digital linguistics with the ASiT project (Sect. 5) and to the access and re-use of scientific data with LoD DIRECT (Sect. 6).

2 Foundations: The DELOS Model and the 5S: Interoperability

The evolution of *Digital Library (DL)* has been favoured by the development of two foundational models of what DL are, namely the *Streams*, *Structures*, *Spaces*,

Scenarios, Societies (5S) model [20] and the DELOS Reference Model [15], which made it clear what kind of entities should be involved in a DL, what their functionalities should be and how Digital Library System (DLS) components should behave, and fostered the design and development of operational DLS complying with them.

However, these two models are quite abstract and, while still providing a unifying vision of what a DL is, they allow for very different choices when it comes to developing actual DLS. This has led to the growth of “ecosystems” where services and components may be able, at best, to interoperate together within the boundaries of DLS that have been inspired by just one of the two models for DL.

In [9] we addressed the need for interoperability among DLS at a high level of abstraction and we showed how this is achieved by a semantically-enabled representation of foundational DL models. The ultimate goal has been to promote and facilitate a better convergence and integration in the context of libraries, archives and museums by lowering the barriers between them.

We proposed a common ontology which encompasses all the concepts considered by the two foundational models and creates explicit connections between

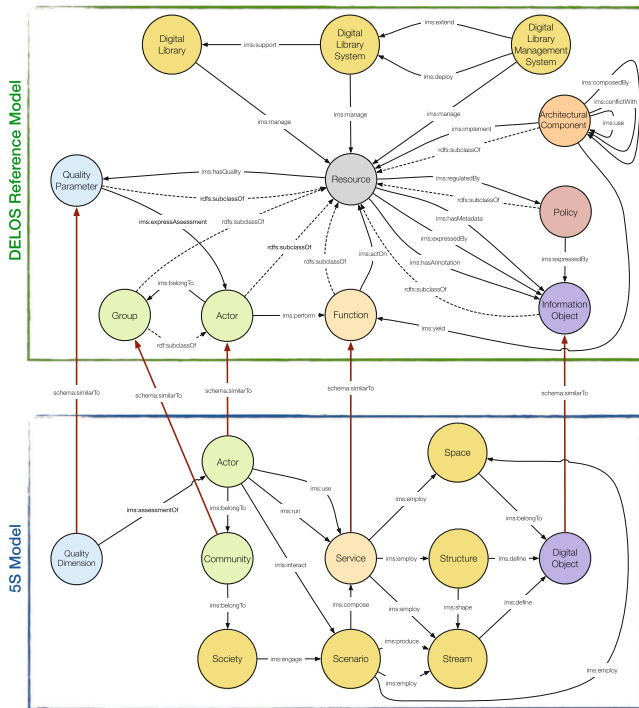


Fig. 1. Semantic mapping of the high level concepts in the 5S model and DELOS Reference Model and their relationships [9].

their constituent domains. In particular, the user, functionality and content domains allow us to enable a high-level interoperability between the actors and the information/digital objects of DL as well as their functions/services.

The DELOS Reference Model and the 5S Model are defined starting from two different viewpoints. Indeed, in DELOS the approach is top-down since it defines the entities and relationships involved in a DL; whereas the 5S model is largely bottom-up starting with key definitions and elucidation of digital library concepts from a minimalist approach. For this reason, some of the concepts modelled by the DELOS Reference Model are not explicitly modelled by the 5S model. The common ontology we defined is particularly effective since it enriches the 5S model with the concepts defined by the DELOS Reference Model, creating further bridges between them and their implementations.

In Fig. 1 we present the *Resource Description Framework (RDF)* graph of the unifying data model relating the DELOS Reference Model to the 5S model by means of a mapping between their most relevant high-level concepts. The presented RDF graph is a visual overview of the ontology we developed.

3 Digital Library Research Infrastructures: The CULTURA Environment

The CULTURA environment is service oriented and is composed of a set of services which integrate to create a rich and engaging experience that supports users of different categories which range from academic and professional users to the general public. The services are conceived and developed to be applicable to a wide variety of cultural collections. The potential generality of the environment is demonstrated by the fact that CULTURA is supporting different use cases that are represented by the *Imaginum Patavinae Scientiae Archivum (IPSA)* and 1641 collections, which differ in morphology, language, modality and metadata. This means that the environment and the supported services need to consider the peculiarities of different documents and different ways of making use of them by diverse categories of users. One of the supported services which must be conceived and made available, taking into specific account the peculiarities of the documents of different collections, is the annotation service [3].

Almost everybody is familiar with annotations and has his own intuitive idea about what they are, drawn from personal experience and the habit of dealing with some kind of annotation in everyday life, which ranges from jottings for the shopping to taking notes during a lecture or even adding a commentary to a text. This intuitiveness makes annotations especially appealing for both researchers and users: the former propose annotations as an easily understandable way of performing user tasks, while the latter feel annotations to be a familiar tool for carrying out their own tasks. Therefore, annotations have been adopted in a variety of different contexts, such as content enrichment, data curation, collaborative and learning applications, and social networks, as well as in various information management systems, such as the Web (semantic and not), digital libraries, and databases.

The role of annotations in digital humanities is well known and documented [2,5,7]. Subsequently, many different tools which allow for the annotation of digital humanities content have been developed. Unfortunately, tools designed specifically for an individual portal are typically only compatible with that system. More general solutions, which can be easily distributed across various sites, have been developed, but these systems often have limited functionality (only annotating a single content type, no sharing features etc.). FAST-CAT (Flexible Annotation Semantic Tool - Content Annotation Tool) is a generic annotation system that directly addresses this challenge by providing a convenient and powerful means of annotating digital content. Figure 2 shows an example of an annotation supported by the CULTURA environment. According to this model, an annotation is a compound multimedia object which is constituted by different signs of annotation. Each sign materializes part of the annotation itself; for example, we can have textual signs, which contain the textual content of the annotation, image signs, if the annotation is made up of images, and so on. In turn, each sign is characterized by one or more meanings of annotation, which specify the semantics of the sign; for example, we can have a sign whose meaning corresponds to the title field in the Dublin Core (DC) metadata schema, in the case of a metadata annotation, or we can have a sign carrying a question of the authors about a document whose meaning may be “question” or similar. An annotation has a scope which defines its visibility (public, shared, or private), and can be shared with different groups of users. Public annotations can be read by everyone and modified only by their owner; shared annotations can be modified by their owner and accessed by the specified list of groups with the given access permissions, e.g., read only or read/write; private annotations can be read and modified only by their owner.

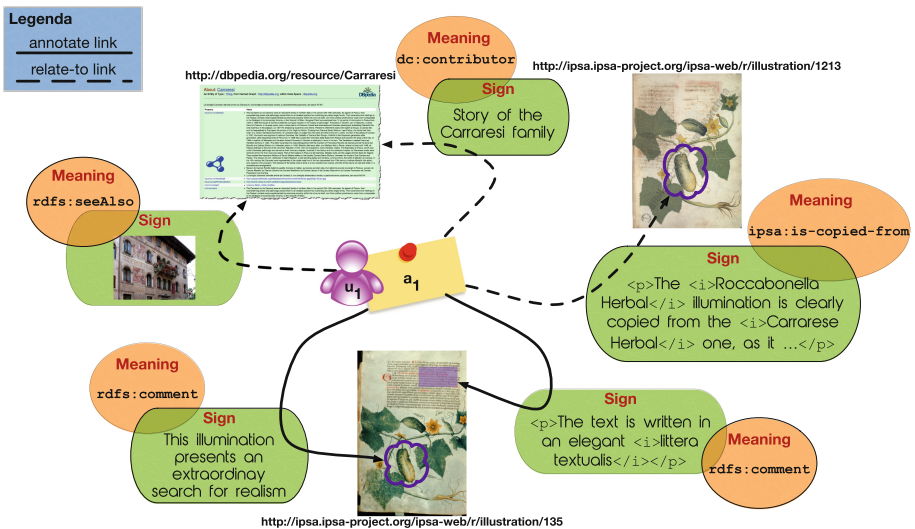


Fig. 2. Example of annotation.

4 Digital Archives: SIAR

The main characteristics of archives are their structure and the objects they manage and preserve. An archive is a complex organization composed by several parts. The foremost component regards the descriptive part of an archive which is conceptually modelled by the *International Standard for Archival Description (General)* (*ISAD(G)*) standard defining the hierarchical organization of archival descriptions and how to model the relationships between them.

We point out two main aspects that we have to consider when modelling an archive: *hierarchy* and *context*. The first aspect means that we have to be able to represent and maintain the hierarchical structure of an archive and its descriptions; the second aspect means that we have to retain the relationships between the archival descriptions and to exploit them to reconstruct the context of a document in relationship with its creation and preservation environment. In order to express hierarchy and context we need a model which allows us to represent the structure of an archive. Furthermore, we also need to represent the content of an archive which is described and managed by means of archival descriptions – that in a digital environment are represented by archival metadata.

SIAR (*Sistema Informativo Archivistico Regionale*) was a project supported by the Italian Veneto Region, the aim of which was to design and develop a digital archive system. The main goal of the SIAR project was to develop a system for managing and sharing archival metadata in a distributed environment and to allow archivists to describe archival material in a collaborative fashion [8].

The context of the work is defined by a group of archivists working in the territory and centrally coordinated by a management and control office of the Veneto Region. The main task of the archivists is to describe the archives of pertinence and produce four main elements: an archival tree organizing the archival descriptive metadata, the descriptions of the preserver, the description of the producer and the finding aids.

The architecture of the system consists of three layers – data, application, and interface logic layers – in order to achieve a better modularity and to properly describe the behaviour of the service by isolating specific functionalities at the proper layer.

The SIAR system is exposed as a RESTful Web Service which allows us to develop different applications and plug-ins over it in an open, collaborative and scalable way which ensure sustainability over time.

The architecture of SIAR is designed at a high level of abstraction in terms of abstract *Application Program Interface (API)* using an object-oriented approach. In this way, we can model the behaviour and the functioning of SIAR without worrying about the actual implementation of each component. Different alternative implementations of each component can be provided, still keeping a coherent view of the whole architecture of the SIAR system.

We achieve this abstraction level by means of a set of interfaces, which define the behaviour of each component of SIAR in abstract terms. Then, a set of abstract classes partially implement the interfaces in order to define the actual behaviour common to all of the implementations of each component. Finally,

the actual implementation is left to the concrete classes, inherited from the abstract ones, that fit SIAR into a given architecture. Furthermore, we apply the abstract factory design pattern, which uses a factory class that provides concrete implementations of a component, compliant with its interface, in order to guarantee a consistent way of managing the different implementations of each component.

Finally, the presentation logic and part of the business logic are implemented via a Liferay Web application, which manages the interaction with the user, controls the flow of the application and translates it into proper *Asynchronous JavaScript Technology and XML (AJAX)* calls to the SIAR RESTful Web Service.

At the core of the system there is the *NEsted SeTs for Object hieRarchies (NESTOR)* model [18, 19], which is composed of two set data models called *Nested Set Model (NS-M)* and *Inverse Nested Set Model (INS-M)*; these two set data models allow us to model hierarchically structured resources by means of an organization of nested sets that is particularly well-suited to archives. The set data models are independent from the tree but they are strongly related to it. Together with the archivists we discussed these data models, pointing out that if we apply them to the archives we are able to maintain the hierarchical structure and the context as well as we can do with the tree data structure, but at the same time they granted us new possibilities of overcoming some of the issues that were highlighted in the ideation phase.

The SIAR system is currently used by the archivists of the Veneto Region for describing and accessing the publicly available archival material of the territory and it is available at the following URL: <http://siar.regione.veneto.it/>.

5 Digital Linguistics: ASiT

Language Resources (LRs) are very important in the development of applications for overcoming language barriers, documenting endangered languages, and for supporting research of several fields. Given the impact of LR, the methodological and technological boundaries existing in linguistic projects need to be overcome in order to find common grounds where linguistic material can be shared and re-used over a long period of time. Consequently, a possibly standardized methodology for designing linguistic databases is necessary to develop linguistic resources that fully meet the desiderata of The FLaReNet Strategic Agenda which presented a set of recommendations for the development and progress of LR in Europe [24].

One of the basic problems we have to deal with when setting up a database with linguistic data is related to the qualitatively and quantitatively different types of data that have to be classified and retrieved. A linguistic database with the function of the old linguistic atlases (and hopefully many more) contains in addition to the obvious linguistic data also many other kinds of data among those information about geographic locations, the type of inquiries adopted to gather the data, the speakers who have delivered the data, all of them being relevant to the linguistic analysis and therefore to be made accessible to the user.

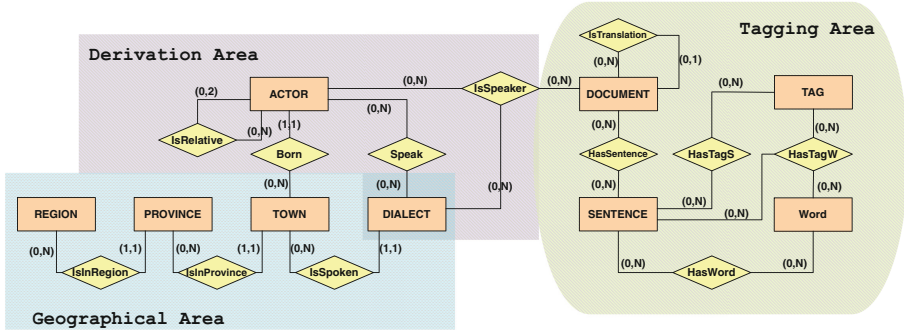


Fig. 3. Entity-relationship diagram for the ASIt Digital Library.

The ASIt (Syntactic Atlas of Italy) linguistic project builds on a long standing tradition of collecting and analyzing linguistic corpora, which has given rise to different studies and projects over the years [1, 10–12]. Research on the syntax of Italian is of great interest to several important lines of research in linguistics: it allows comparison between closely related varieties (the dialects), hence the formation of hypotheses about the nature of cross-linguistic parametrization; it allows contact phenomena between Romance and Germanic varieties to be singled out, in those areas where Germanic dialects are spoken; it allows syntactic phenomena of Romance and Germanic dialects to be found, described and analyzed to a great level of detail [10]. The conceptual model of ASIt is depicted in Fig. 3.

The ASIt Digital Library System was originally intended to support the first line of research, i.e. comparison between closely related Italian varieties [1]. The corpus to be automatically handled was firstly envisioned and secondly mapped on a conceptual schema in order to be general enough to handle diversified geolinguistic projects with tagging on different linguistic units. This was a crucial methodological investment to support the other lines of research, specifically those involving the relationship between Romance and Germanic varieties and investigated in a multidisciplinary and collaborative project, “Cimbrian as a test case for synchronic and diachronic language variation” [10].

Exposing linguistic data as Linked Open Data enhances the interoperability between existing linguistic datasets and allows for their integration with other resources that use a Resource Description Framework (RDF) approach such as lexical-semantic resources already available as Linked Data, e.g. a general knowledge base like DBpedia¹, or linguistic resources like WordNet² or Wiktionary³. In order to make ASIt re-usable and interoperable, we defined the ASIt Linguistic Linked Dataset based on the conceptual schema of the curated database [14]. In Fig. 4 we report the main classes and properties defining the RDF schema.

¹ <http://wiki.dbpedia.org>.

² <https://wordnet.princeton.edu>.

³ <https://www.wiktionary.org>.

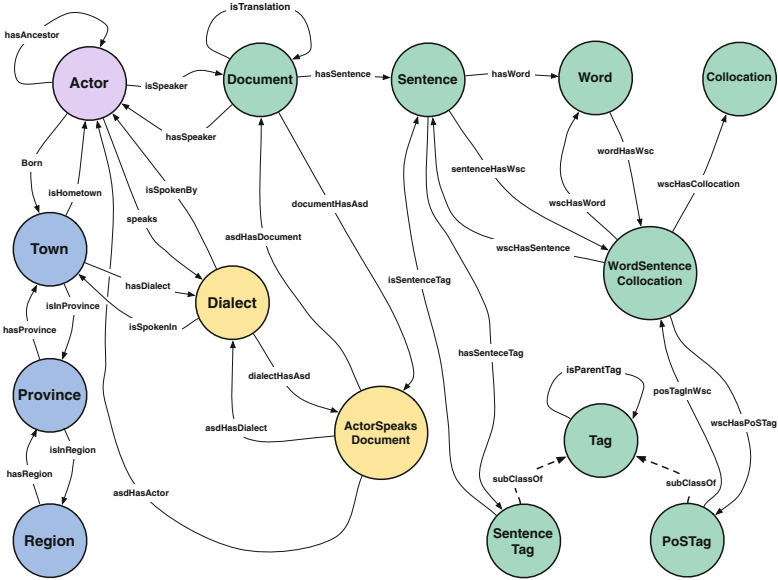


Fig. 4. Diagram representing the RDF/S defined for the ASIIt enterprise.

The generalizability of the ASIIt approach that is materialized in the developed conceptual schema has been shown in a recent test case [11]: the DFG-Projekt PO 1642/1-1.⁴ The objective of this project is the synchronic and diachronic analysis of the syntax of Italian and Portuguese relative clauses. Since the project aimed at investigating a set of phenomena related to different types of relative clauses, syntactic phenomena under investigation are captured through a new dedicated sentence level tag set tailored for this project. This database is the first attempt to investigate different types of relative clauses in a corpus of spoken colloquial language in a systematic way. The challenge consisted in adapting the tools of the ASIIt project to the corpus data, i.e. adapting a design originally created to deal with a purely experimental setting to a much freer and less controlled set of data coming from a pre-existing corpus.

6 Scientific Data: The LoD DIRECT System

The importance of research data is widely recognized across all scientific fields as this data constitutes a fundamental building block of science. Recently, a great deal of attention was dedicated to the nature of research data [13] 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 [16, 17, 23].

⁴ <http://ims.dei.unipd.it/websites/portuguese-relclauses/index.html>.

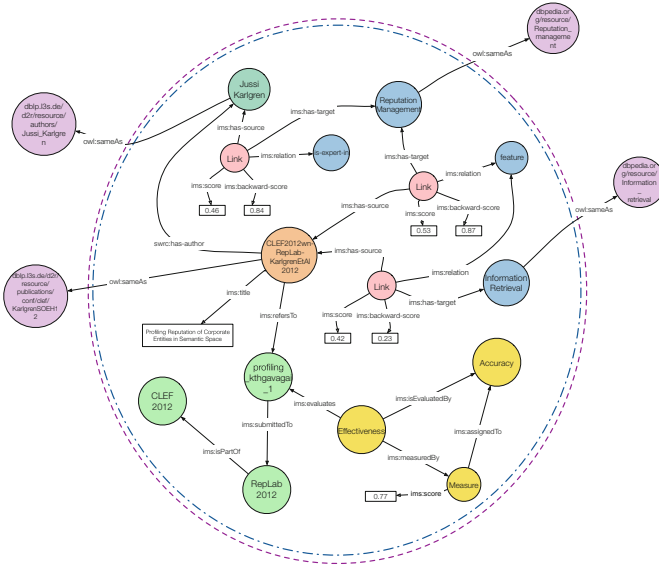


Fig. 5. An example of RDF graph showing how expertise topics and expert profiles are used for enriching IR experimental data.

Nevertheless, in the field of *Information Retrieval (IR)*, where experimental evaluation based on shared data collections and experiments has always been central to the advancement of the field [21], the *Linked Open Data (LOD)* paradigm has not been adopted yet and no models or common ontologies for data sharing have been proposed. So despite the importance of data to IR, the field does not share any clear ways of exposing, enriching, and re-using experimental data as LOD with the research community.

We discuss an example of the outcomes of the semantic modelling and automatic enrichment processes applied to the use case of discovering, understanding and re-using the experimental data; the details of the full RDF model are reported in [22]. Figure 5 shows an RDF graph, which provides a visual representation of how the experimental data are enriched. In particular, we can see the relationship between a contribution and an author enriched by expertise topics, expert profiles and connections to the LOD cloud, as supported by the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* system which provides the conceptual model for representing and enriching the data [4, 6].

In this instance, the author (*Jussi Karlgren*) and the contribution (*KarlgrenEtAl-CLEF2012*) are data derived from the evaluation workflow, whereas all the other information are automatically determined by the enrichment process. The adopted methodology for expertise topics extraction determined two main topics, “reputation management” and “information 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 (normalized in the interval [0, 1]) give a measure of how much this contribution is about a specific topic and we can see that in this case it is concerned a bit more with reputation management than with information retrieval. Furthermore, the backward-score gives us additional information by measuring how much a contribution is authoritative with respect to a scientific topic. In Fig. 5, we can see that *KarlgrenEtAl-CLEF2012* is authoritative for reputation management (backward-score of 0.87), whereas it is not a very important reference for information retrieval (backward-score of 0.23). Summing up, we can say that if we consider the relation between a contribution and an expertise topic, the score indicates the pertinence of the expertise topic within the contribution; whereas the backward score indicates the pertinence of the contribution within the expertise topic. The higher the backward score, the more pertinent is the contribution for the given topic.

This information is confirmed by the expert profile data; indeed, looking at the upper-left part of Fig. 5, the author *Jussi Karlgren* is considered “an expert in” reputation management (backward-score of 0.84), even if it is not his main field of expertise (score of 0.46).

All of this automatically extracted information enriches the experimental data enabling for a higher degree of re-usability and understandability of the data themselves. In this use case, we can see that the expertise topics are connected via an `owl:sameAs` property to external resources belonging to the DBpedia⁵ linked open dataset. These connections are automatically defined via the semantic grounding methodology described below and enable the experimental data to be easily discovered on the Web. In the same way, authors and contributions are connected to the DBLP⁶ linked open dataset.

In Fig. 5 we can see how the contribution (*KarlgrenEtAl-CLEF2012*) is related to the experiment (*profiling_kthgavagai_1*) on which it is based. This experiment was submitted to the *RepLab 2012* of the evaluation campaign *CLEF 2012*. It is worthwhile to highlight that each evaluation campaign in DIRECT is defined by the name of the campaign (CLEF) and the year it took place (e.g., 2012 in this instance); each evaluation campaign is composed of one or more tasks identified by a name (e.g., RepLab 2012) and the experiments are treated as submissions to the tasks. Each experiment is described by a contribution which reports the main information about the research group which conducted the experiment, the system they adopted, developed and any other useful detail about the experiment.

We can see that most of the reported information is directly related to the contribution and they allow us to explicitly connect the research data with the scientific publications based on them. Furthermore, the experiment is evaluated from the “effectiveness” point of view by using the “accuracy” measurement which has 0.77 score. Retaining and exposing this information as LOD on the

⁵ <http://www.dbpedia.org/>.

⁶ <http://dblp.l3s.de/>.

Web allow us to explicitly connect the results of the evaluation activities to the claims reported by the contributions.

The described RDF model has been realized by the DIRECT [4,6] system which allows for accessing the experimental evaluation data enriched by the expert profiles created by means of the techniques that will be described in the next sections. This system is called LOD-DIRECT and it is available at the URL: <http://lod-direct.dei.unipd.it/>.

The data currently available include the contributions produced by the *Conference and Labs of the Evaluation Forum (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 expertise topics which are available as linked data as well.

At the time of writing, LOD-DIRECT allows access to 2,229 contributions, 2,334 author profiles and 2,120 expertise topics. Overall, 1,659 experts have been individuated and on average there are 8 experts per expertise topics (an expert can have more than one expertise of course).

References

1. Agosti, M., Benincà, P., Di Nunzio, G.M., Miotto, R., Pescarini, D.: A digital library effort to support the building of grammatical resources for Italian dialects. In: Agosti, M., Esposito, F., Thanos, C. (eds.) IRCDL 2010. CCIS, vol. 91, pp. 89–100. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15850-6_10
2. Agosti, M., Bonfiglio-Dosio, G., Ferro, N.: A historical and contemporary study on annotations to derive key features for systems design. *Int. J. Digital Libr. (IJDL)* **8**(1), 1–19 (2007)
3. Agosti, M., Conlan, O., Ferro, N., Hampson, C., Munnely, G.: Interacting with digital cultural heritage collections via annotations: the CULTURA approach. In: Marinai, S., Marriot, K. (eds.) *Proceedings of 13th ACM Symposium on Document Engineering (DocEng 2013)*, pp. 13–22. ACM Press (2013)
4. Agosti, M., Di Buccio, E., Ferro, N., Masiero, I., Peruzzo, S., Silvello, G.: DIRECTIONS: design and specification of an IR evaluation infrastructure. In: Catarci, T., Forner, P., Hiemstra, D., Peñas, A., Santucci, G. (eds.) *CLEF 2012*. LNCS, vol. 7488, pp. 88–99. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33247-0_11
5. Agosti, M., Ferro, N.: A formal model of annotations of digital content. *ACM Trans. Inf. Syst. (TOIS)* **26**(1), 3:1–3:57 (2008)
6. Agosti, M., Ferro, N.: Towards an evaluation infrastructure for DL performance evaluation. In: *Evaluation of Digital Libraries: An Insight into Useful Applications and Methods*, pp. 93–120. Chandos Publishing, Oxford (2009)
7. Agosti, M., Ferro, N., Frommholz, I., Thiel, U.: Annotations in digital libraries and collaboratories – facets, models and usage. In: Heery, R., Lyon, L. (eds.) *ECDL 2004*. LNCS, vol. 3232, pp. 244–255. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30230-8_23

8. Agosti, M., Ferro, N., Rigon, A., Silvello, G., Terenzoni, E., Tommasi, C.: SIAR: a user-centric digital archive system. In: Agosti, M., Esposito, F., Meghini, C., Orio, N. (eds.) IRCDL 2011. CCIS, vol. 249, pp. 87–99. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-27302-5_8
9. Agosti, M., Ferro, N., Silvello, G.: Digital library interoperability at high level of abstraction. *Future Gener. Comput. Syst.* **55**, 129–146 (2016)
10. Agosti, M., Alber, B., Di Nunzio, G.M., Dussin, M., Rabanus, S., Tomaselli, A.: A curated database for linguistic research: the test case of Cimbrian varieties. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012). European Language Resources Association (ELRA) (2012)
11. Agosti, M., Di Buccio, E., Di Nunzio, G.M., Poletto, C., Rinke, E.: Designing a long lasting linguistic project: the case study of ASIT. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016 (2016)
12. Benincà, P., Poletto, C.: The ASIS enterprise: a view on the construction of a syntactic atlas for the Northern Italian dialects. *Nordlyd* **34**, 35–52 (2007). Monographic issue on Scandinavian Dialects Syntax
13. Borgman, C.L.: *Big Data, Little Data, No Data*. MIT Press, Cambridge (2015)
14. Di Buccio, E., Di Nunzio, G.M., Silvello, G.: A curated and evolving linguistic linked dataset. *Semant. Web* **4**(3), 265–270 (2013)
15. 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.: The DELOS digital library reference model. Foundations for Digital Libraries. ISTI-CNR at Gruppo ALI, Pisa, Italy (2007)
16. Ferro, N.: Reproducibility challenges in information retrieval evaluation. *ACM J. Data Inf. Qual. (JDIQ)* **8**(2), 8:1–8:4 (2017)
17. Ferro, N., Fuhr, N., Järvelin, K., Kando, N., Lippold, M., Zobel, J.: Increasing reproducibility in IR: findings from the Dagstuhl Seminar on “Reproducibility of data-oriented experiments in e-Science. *SIGIR Forum* **50**(1), 68–82 (2016)
18. Ferro, N., Silvello, G.: NESTOR: a formal model for digital archives. *Inf. Process. Manage.* **49**(6), 1206–1240 (2013)
19. Ferro, N., Silvello, G.: Descendants, ancestors, children and parent: a set-based approach to efficiently address XPath primitives. *Inf. Process. Manage.* **52**(3), 399–429 (2016)
20. Gonçalves, M.A., Fox, E.A., Watson, L.T., Kipp, N.A.: Streams, structures, spaces, scenarios, societies (5S): a formal model for digital libraries. *ACM Trans. Inf. Syst. (TOIS)* **22**(2), 270–312 (2004)
21. Harman, D.K.: *Information Retrieval Evaluation*. Morgan & Claypool Publishers, San Raphael (2011)
22. Silvello, G., Bordea, G., Ferro, N., Buitelaar, P., Bogers, T.: Semantic representation and enrichment of Information Retrieval experimental data. *Int. J. Digital Libr. (IJDL)* **18**(2), 145–172 (2017)
23. Silvello, G., Ferro, N.: Data citation is coming, Introduction to the special issue on data citation. *Bull. IEEE Tech. Committee Digital Libr. (IEEE-TCDL)*, **12**(1), 1–5 (2016)
24. Soria, C., Calzolari, N., Monachini, M., Quochi, V., Bel, N., Choukri, K., Mariani, J., Odijk, J., Piperidis, S.: The language resource strategic agenda: the flaret synthesis of community recommendations. *Lang. Resour. Eval.* **48**(4), 753–775 (2014)

Thirty Years of Digital Libraries Research at the University of Padua: The User Side

Maristella Agosti, Giorgio Maria Di Nunzio^(✉), Nicola Ferro, Maria Maistro, Stefano Marchesin, Nicola Orio, Chiara Ponchia, and Gianmaria Silvello

University of Padua, Padua, Italy

{maristella.agosti,giorgiomaria.dinunzio,nicola.ferro,maria.maistro, stefano.marchesin,nicola.orio,chiara.ponchia,gianmaria.silvello}@unipd.it

Abstract. For the 30th anniversary of the Information Management Systems (IMS) research group of the University of Padua, we report the main and more recent contributions of the group that focus on the users in the field of Digital Library (DL). In particular, we describe a dynamic and adaptive environment for user engagement with cultural heritage collections, the role of log analysis for studying the interaction between users and DL, and how to model user behaviour.

1 Introduction

Traditionally, Digital Libraries (DL) are considered places where information resources can be stored and made available to end users, but in the last decade they have also become systems that need to support the user in different information centric activities. Indeed, in the context of DL we need to take into account heterogeneous information sources with different community background such as libraries, archives and museums; but, DL are central also for research purposes and they provide the infrastructures able to gather, manage and grant access to scientific data at large. When it comes to interacting with the system for discovering, retrieving, re-using or citing cultural heritage objects or scientific data, users are the main focus for DL systems.

In the last thirty years, the Information Management Systems (IMS) research group at the University of Padua has contributed substantially to the DL field with special attention to the role of users. In this paper, we report the more recent and relevant contributions that focus (i) on user engagement with cultural heritage collections in the context of the CULTURA project (Sect. 2), (ii) on the role of log data analysis in the context of multilingual DL and search engines (Sect. 3), and (iii) on user behaviour modelling (Sect. 4).

2 Features of Adaptivity in the CULTURA Environment

The main goal of CULTURA, a European project co-funded under the 7th Framework Programme which ran from 2011 to 2014¹, was to increase user

¹ <http://www.cultura-strep.eu/>.

engagement with digital cultural heritage collections through the development of a new adaptive and dynamic environment, that is a Virtual Research Environment (VRE), and specifically developed tools. The CULTURA consortium had a strong emphasis on meeting real end-user needs, maximizing societal impact and laying a foundation for successful commercialization. To this end, the environment went beyond the traditional search-based exploration, providing natural language processing technologies, entity-oriented search and a comprehensive set of logging, bookmarking and annotating tools that make it a powerful aid to both extensive and intensive work on content collections. For the validation of the project, two pre-existing cultural heritage collections were used: The *1641 Depositions*, a collection of accounts by victims of the Irish Rebellion of 1641, which constitutes a textual corpus that has been augmented by manually generated metadata²; and *IPSA*, a collection of illuminated scientific manuscripts (including herbals and astronomical-astrological codices), which is a purely visual collection with extensive metadata³.

2.1 Narratives

Narratives have been introduced in the CULTURA environment as a novel tool to engage the different types of users [10]. Narratives are implemented as threads through the document collection, linking artifacts and tools related to a particular topic. Expert researchers, guided by the use cases and user requirements outlined during user consultations, designed narratives. They used their specialist knowledge of the collections to create a series of paths through the content that can engage users from all the groups in the exploration and use of specific content. To address the different level of expertise of each category of users, each narrative has a number of levels. Less expert users are offered a relatively high level narrative, but as users interact with the resources that are presented to them, the system dynamically discloses additional material, resulting in a more complex and more captivating user experience. These narratives allow for an open-ended and developing engagement with the resource collections. An example of a high level narrative are the ones designed for less expert users. These describe the different steps of a short course on a specific topic that is encompassed within the collection. Typically, the relevant material will be spread across the diverse parts of the different components of a collection. Adaptive narratives provide structured routes through the collection, exposing the user to artifacts that are relevant to their topic of interest. Furthermore, at certain steps of the narrative, users are given the possibility to access some extra steps which will provide them with additional information on that particular topic, also using external resources such as Wikipedia. Once they have gone through all the additional steps, users will come back to the point where they left the narrative and will be able to proceed.

These narratives were implemented for both collections in the environment, and the user experience of these narratives validated their usefulness for both

² <http://1641.tcd.ie/>.

³ <http://ipsa.dei.unipd.it/>.

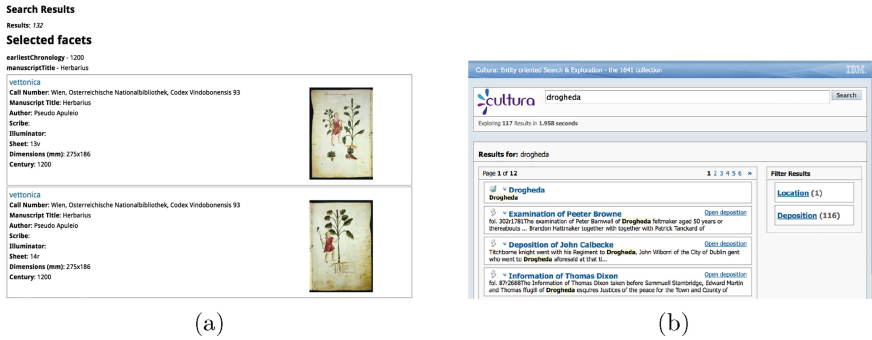


Fig. 1. Faceted search and entity-oriented search.

types of content collection. In the case of the historical textual corpus, a series of narratives were developed for use with secondary school students who were encountering the collection for the first time. These users were presented with a sidebar containing a brief explanation of the context of the individual part of the text being viewed, along with prompts for further research. Users can move backwards or forwards within the narrative, or branch off into further narratives, covering particular areas that especially pique their interest. At any point, the user can leave the pathway to carry out their own detailed investigations. Users are able to resume the narrative when desired. In addition to tying together chains of documents, these narratives can include the other tools that contribute to the environment. For instance, a user of the IPSA collection who is following a narrative based on herbals of a given historical period can be presented with the results of a faceted search based on chronology and title (see Fig. 1a) or a user who is following a narrative based on a particular location can be presented with the results of an entity-oriented search displaying the relationship of the entity in question to other entities within the collection (see Fig. 1b). The visualization features are also available to users of the textual collection, and, in a similar way, allow the collection to be explored on the basis of the links and relationships between people, places and events. Individually, these tools offer a range of exciting new ways of looking at the collection.

The narratives add a further layer of richness, integrating these views within a coherent structure. Narratives help in fostering a more involving experience of the collections, more complete and extended through time. Therefore, other kinds of static collections can benefit from narratives too, e.g. archaeological collections.

2.2 An Adaptive Cross-Site User Modelling Platform for Cultural Heritage Websites

Websites need information about their users in order to deliver personalized experiences [20, 23]. Previous activity across the Web provides an opportunity

to gather further data about the user from beyond the website, enhancing user profiles with overarching information from different websites. In cross-site user modelling, web personalization techniques are based upon information provided by an independent third-party user modelling service, in order to assist users in addressing information needs.

In the domain of cultural heritage, this approach can provide relevant information to users that attempt to answer cross-site information needs within their browsing space. In those situations where the user's need spans topics that are not confined to a single website, the introduction of such cross-site service [24] might improve the effectiveness of the website personalization, along with the user's level of satisfaction. In [9], a parallel and complementary approach to the Virtual Research Environment (VRE) for the Digital Humanities of the CULTURA (see footnote 1) project [32,33] has been proposed. Introducing the cross-site service, the user model can gain a higher precision in those topics that are more relevant to the user and provide more tailored personalization to help addressing the user's cross-site information needs. An example of this could be an overarching user's interest in the living conditions of the Irish middle class during the Irish rebellion, which cannot be addressed by a single website of the CULTURA VRE but requires the user to navigate across different websites of the VRE browsing space.

The general process that identifies the cross-site approach is as follows: (1) The user lands on a website related to an information gathering task in the cultural heritage domain; (2) The user authenticates a first time with the third party service; (3) The website tracks all user activities in the webpages along with the relevant text entities identified by a term identification component; (4) The user triggers the information exchange function, in anticipation of subsequent personalisation by the target website, which provides relevant user data (based on the selected communication pattern) to the website and newly tracked user information to the service; (5) The user surfs to a second website and, depending on whether they are already authenticated or not, authenticates or directly triggers the information exchange function, which should provide more tailored information to the website; (6) Steps (1)–(5) can then be re-iterated many more times, without a strict order of execution. The high-level architecture of the cross-site service is presented below and can be found in Fig. 2.

Term Identification Component: The main purpose of the term identification component is to identify text entities related to the body of the current webpage the user is viewing. Text entities indicate the meaning of the underlying content from websites belonging to the cross-site browsing space. An additional responsibility of the term identification component is to ensure the creation of a shared conceptualization of the user's cross-site browsing space. This conceptualization is represented as a text entity space and it is based on the contents the user has browsed within the websites of the cross-site browsing space.

User Profile Component: The user profile consists of activities and text entities related to the current task of the user and that were identified by the term identification component.

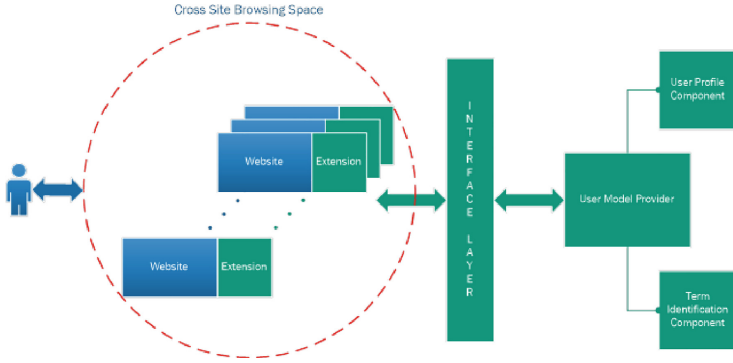


Fig. 2. High-level service architecture.

Interface Layer: The interface layer provides an abstraction from the specific implementation of the different websites within the user’s cross-site browsing space. It implements a RESTful API⁴ that facilitates the communication between the cross-site service and the websites the user is browsing.

WCMS - Web-Based Content Management System Module Extensions: WCMS module extensions allow a simple and limited-impact integration of cross-site information exchange techniques into existing website implementation. The responsibility of the WCMS module extensions is twofold: (1) To facilitate communication between the website and the API of the cross-site user modelling platform and (2) to provide non-intrusive information exchange techniques to the user, within the website the user is currently browsing.

Cross-Site Browsing Space: The application of information exchange techniques to independently hosted websites introduces a cross-site browsing space. Within the browsing space, target websites receive user data through information exchange techniques and return novel user’s data to the cross-site service through the same mechanism.

3 Log Analysis

The interaction between users and information access systems can be analyzed and studied for different goals, for example to personalize the presentation of results. User preferences can be learned either explicitly with surveys or questionnaires, or implicitly by studying the actions that the user performs when using the system. These actions are saved on log files that can be used to study the usage of a specific application, and to better adapt it to the objectives the users were expecting to reach. The analysis of log data can be roughly divided according to the nature of the system, as suggested by [4], into: Web search engines (WSE) log analysis and Digital Library Systems (DLS) log analysis.

⁴ http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch.style.htm.

In fact, both WSE and DLS can be used as information access systems accessible through the Web; however, the collections of documents the user can access are drastically different: WSE retrieve documents that are in general Web pages, while DLS retrieve documents that have been chosen after a quality control performed by professionals.

3.1 The European Library (TEL)

The European Library⁵ originates from the TEL project that was partly founded by the European Commission and successfully ended on January 2004. A new project called TEL-ME-MOR⁶ was born with the purpose to expand The European Library in order to support the ten national libraries from the New Member States of the European Union and later to offer access to the resources of 45 European National Libraries.

In [7], a general methodology for gathering and mining information from Web log files was proposed and a series of tools to retrieve, store, and analyze the data extracted from Web logs was designed and implemented. The aim of this work was to form general methods by abstracting from the analysis of the Web logs of TEL in order to give advice about the development of the portal from the point of view of both the security and the improvement of personalization. The solution proposed by this methodology for the problem of storing the information contained in a log file was the use of a DataBase Management System (DBMS) in such a way that it was possible to perform queries easily, for example to obtain statistical information useful for the development of the site and to allow subsequent mining of the managed data. Figure 3 presents the Entity-Relationship conceptual schema that was designed.

In [8], an experimental analysis was performed on the available log data at that time, corresponding to eleven months of TEL Web log files from October 31st 2005 to September 25th 2006. In this analysis, we used heuristics to identify users and, as a result, we suggested that authentication would be required since it would allow TEL servers to identify users and create profiles to tailor specific needs. Moreover, authentication would have also helped to solve the problem concerning crawlers accesses, granting access to some sections of the Web site only to registered users, blocking crawlers using faked user agents. A second batch of seven months of TEL Web log files, from October 1st 2006 to April 30th 2007, was analyzed in [1]. This new set of log files contained richer information about: the Internet Protocol (IP) address and the user-agent which allowed the identification of single users, as well as the referrer field, a Uniform Resource Locator (URL) address which communicates the last page viewed by the user which could be used to know the way visitors get to TEL service. These logs also contained the information saved by the cookie file such as: the language selected by the user during the navigation of the service; the collections of documents selected during the query or query refinement, the identifier of

⁵ <http://www.theeuropeanlibrary.org/>.

⁶ <http://telmemor.net/>.

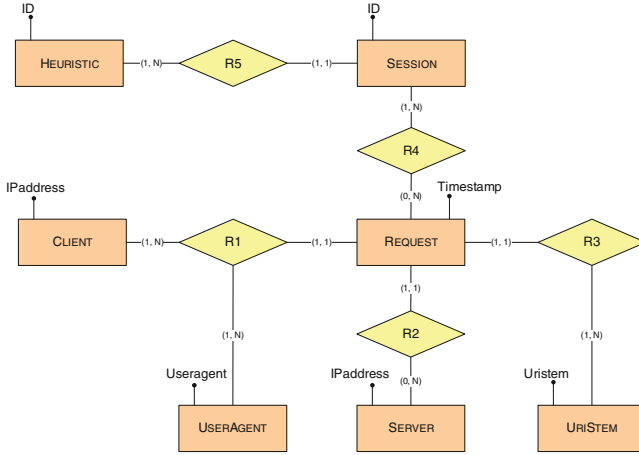


Fig. 3. The European Library Log Data Storage Entity-Relationship Schema. Rectangles show the main objects (entities) of the data contained in the log files. Diamond shapes describe the relationships among the objects.

the session assigned by the server to a specific user. These richer data allowed us to better understand the profile of the TEL users in terms of geographical areas, willingness to spend some time to perform advanced searches and to select collections of documents different from the default ones. Further analysis to combine the results of the HTTP logs analysis with that of the user study to give a better understanding on the usage of The European Library service were presented in [3, 6]. In this work, a user study was conducted in order to collect into HTTP logs enough data to study the browsing activity and analyze possible relations between explicit preferences collected by online questionnaires and implicit actions recorded in the logs. One interesting finding of this implicit and explicit data analysis was that short sessions found in the HTTP logs could be explained by the answers collected by questionnaires during user studies; in particular, users were not satisfied by the presentation of the results and by the content of the results. The sources of data used and the interrelation among them are depicted in Fig. 4 where an example of a user session is drawn.

Implicit data on user interaction with a new Digital Library portal were useful for detecting typical usage patterns of different user groups, as well as their possible evolution or stability over time. They provided key information for identifying possible interaction problems on a large-scale (e.g. the whole user population), but little or no insight as to why these problems occurred and how they could be solved, as discussed in [5]. They undoubtedly were influential in the design of more focused survey studies (based on explicit user data) able to investigate which relevant user expectations, habits, motivations, preferences or difficulties were responsible for the interaction patterns observed. Some of the insights gathered in these surveys were used as requirements for the design of an advanced query suggestion tool focussed on and tailored to TEL named i-TEL-u.

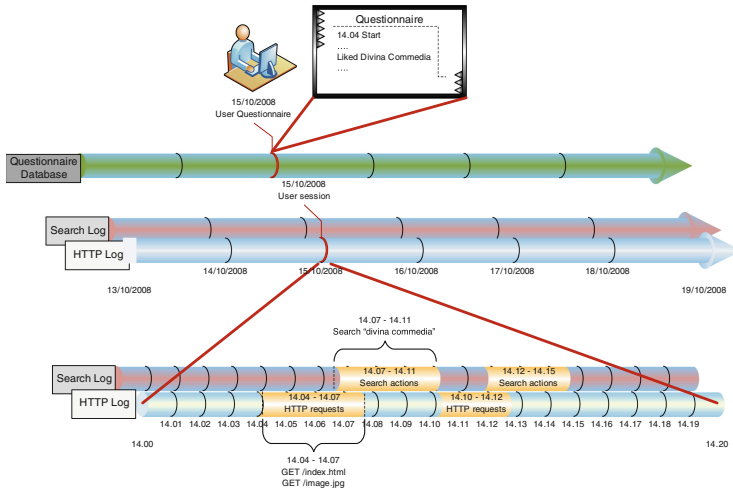


Fig. 4. The three arrows represent the three sources of data that are collected and combined together by the proposed method. An example of a user session is shown to highlight the way the three sources are generated in an interleaved way during the user activities.

While traditional query suggestion tools could not recognize different contexts from which the suggested queries are produced, i-TEL-u leveraged a variety of data sources and allowed users to seamlessly move from one context to another according to their evolving information needs during the search session [2].

3.2 Multilingual Log Analysis - LogCLEF

Despite being a very important resource to study user information needs and preferences, very few log datasets have been made available for research experiments. This lack of log data have made the verifiability and repeatability of experiments very limited since it is difficult to find two researches on the same system (for example on the same search engine or digital library), as well as using the same period of time to make results comparable across different studies. In the context of the Cross-Language Evaluation Forum (CLEF) Initiative⁷, a self-organized body whose main mission is to promote research, innovation, and development of information access systems with an emphasis on multilingual and multimodal information with various levels of structure, LogCLEF was an evaluation initiative for the analysis of queries and other logged activities used as an expression of user behaviour [17,25]. An important long-term aim of the LogCLEF initiative was to stimulate research on user behaviour in multilingual environments and promote standard evaluation collections of log data. From 2009 until 2011, LogCLEF released collections of log data with the aim of verifiability and repeatability of experiments [16,27,28]. In the three years of LogCLEF editions, different data sets have been distributed to the participants:

⁷ <http://www.clef-initiative.eu>.

search engine query and server logs from the Portuguese search engine Tumba⁸ and from the German EduServer⁹; digital library systems query and server logs from The European Library (TEL) and Web search engine query logs of the Chinese search engine Sogou¹⁰.

LogCLEF promoted different tasks such as:

- Language identification task: participants are required to recognize the actual language of the query submitted.
- Query classification: participants are required to annotate each query with a label which represents a category of interest (for example, geographic entities, historical events, people, etc.).
- Success of a query: participants are required to study the trend of the success of a search. Success can be defined in terms of time spent on a page, number of clicked items, actions performed during the browsing of the results;
- Query refinding: when a user clicks an item following a search, and then later clicks on the same item via another search;
- Query refinement: when a user starts with a query and then the following queries in the same session are a generalization, specification, or shift of the original one.

4 User-System Interaction

The huge increase in the number of documents published after WWII called for the design and development of new and fully automatic techniques to handle and search this increasing amount of information content. Web search [15], biomedical search [30], patent retrieval [26], and enterprise search [13], as well as DL, started to rely on Information Retrieval Systems (IRS) to perform automatic indexing and searching. However, with Internet becoming pervasive and accessible to everyone, IRS have begun to exploit more and more complex techniques to satisfy the user information need. Among them, particularly advantageous was the use of click log data to adapt systems to users behaviour.

Nowadays, it is a pretty common practice for commercial search engines, to record user interactions with their interface, such as query keywords, clicked URLs, impressions and clicks timestamps and many others. Click log data offer several advantages [22]: they are easy and inexpensive to collect, they are available in real time and they are user centered, i.e. they directly represent user preferences. However, even if click data have proven to be a valuable resource of implicit feedback, they are biased and noisy and they are intrinsically difficult to interpret [21]. Moreover, the user behaviour is tightly related to the information need and task that she is performing. As an example, consider that solely in Web search user queries can be classified as navigational, when the user searches for a particular website that she has in mind, and Informational, when the user

⁸ <http://www.tumba.pt/>.

⁹ <http://www.eduserver.de/>.

¹⁰ <http://www.sogou.com>.

explores the result page looking for some information on a specific topic [11]. When the scope is widened and heterogeneous tasks are considered, then the differences in user behaviour will be even more pronounced, as presented in [31] where Web, job and talent search are compared. In [31] a thorough analysis of log data is conducted, showing that substantial differences in Web, job and talent search can be detected, when considering the click frequency for each rank position, query frequency and query diversity.

The heterogeneity of tasks, information needs and users require IRS to adjust their output by considering the context in which the search is performed. This can be achieved by modelling the user behaviour in order to remove the noise and bias incorporated in log data. In [18] a new user model based on Markov chain is presented. Differently from other state of the art models [14, 29], the exploitation of Markov chains makes it possible to describe a user who scans the ranked list of documents according to possibly complex paths. Therefore, the model can handle users that go forward and backward, jump from one document to any other in the list and visit the same document multiple times.

Furthermore, the proposed model can be embedded in IRS to account for user interactions. In [19] the Markovian model is used to describe the user dynamic, which is successively integrated in the *Learning to Rank (LtR)* algorithm LAMB-DAMART [12, 34]. The proposed approach stems from the observation that users behave differently depending on the query type, i.e. navigational vs. informational, and two different dynamics are calibrated on a click log dataset to resemble the real user dynamic. The algorithm accounts for the user dynamic as discounts for the objective function, assigning different discounting accordingly to the query category.

Finally, personalized IRS require involvement of the user even during the evaluation process. Indeed, if the system returns a different output depending on the user and the task, even the evaluation measure should consider these features. The Markovian model proposed in [18] defines a new family of evaluation measures, called *Markov Precision (MP)*, which injects user models into precision. MP can be considered as an offline measure, by using predefined transition matrices, or as an online measure, by tuning the transition matrix on click log data. In the latter case, the measure attempts to face the challenge of adapting evaluation to the user behaviour.

References

1. Agosti, M., Angelaki, G., Coppotelli, T., Di Nunzio, G.M.: Analysing HTTP logs of a European DL initiative to maximize usage and usability. In: Goh, D.H.-L., Cao, T.H., Sølvsberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 35–44. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-77094-7_9
2. Agosti, M., Cisco, D., Di Nunzio, G.M., Masiero, I., Melucci, M.: i-TEL-u: a query suggestion tool for integrating heterogeneous contexts in a digital library. In: Lalmas, M., Jose, J., Rauber, A., Sebastiani, F., Frommholz, I. (eds.) ECDL 2010. LNCS, vol. 6273, pp. 397–400. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15464-5_41

3. Agosti, M., Crivellari, F., Di Nunzio, G.M.: Evaluation of digital library services using complementary logs. In: *Proceedings of the Workshop on Understanding the User - Logging and Interpreting User Interactions in Information Search and Retrieval* (2009)
4. Agosti, M., Crivellari, F., Di Nunzio, G.M.: Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. *Data Min. Knowl. Discov.* **24**(3), 663–696 (2012)
5. Agosti, M., Crivellari, F., Di Nunzio, G.M., Gabrielli, S.: Understanding user requirements and preferences for a digital library web portal. *Int. J. Digit. Libr.* **11**(4), 225–238 (2010)
6. Agosti, M., Crivellari, F., Di Nunzio, G.M., Ioannidis, Y.E., Stamatogiannakis, E., Triantafyllidi, M.L., Vayanou, M.: Searching and browsing digital library catalogues: a combined log analysis for the European library. In: *Post-Proceedings of the Fifth Italian Research Conference on Digital Libraries - IRCDL 2009*, pp. 120–135 (2009)
7. Agosti, M., Di Nunzio, G.M.: Gathering and mining information from web log files. In: Thanos, C., Borri, F., Candela, L. (eds.) *DELOS 2007*. LNCS, vol. 4877, pp. 104–113. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-77088-6_10
8. Agosti, M., Di Nunzio, G.M.: Web log mining: a study of user sessions. In: *10th DELOS Thematic Workshop on Personalized Access, Profile Management, and Context Awareness in Digital Libraries* (2007)
9. Agosti, M., Lawless, S., Marchesin, S., Wade, V.: An adaptive cross-site user modelling platform for cultural heritage websites. In: *Digital Libraries and Archives*. CCIS. Springer (2017, in print)
10. Agosti, M., Orio, N., Ponchia, C.: Guided tours across a collection of historical digital images. In: *Proceedings of the Third AIUCD Annual Conference, AIUCD 2014*, pp. 7:1–7:6. ACM Press (2015)
11. Broder, A.: A taxonomy of web search. *SIGIR Forum* **36**(2), 3–10 (2002)
12. Burges, C.J.: From ranknet to lambdarank to lambdamart: an overview. Technical report (2010)
13. Burnett, S., Clarke, S., Davis, M., Edwards, R., Kellett, A.: *Enterprise Search and Retrieval. Unlocking the Organisation's Potential*. Butler Direct Limited, Kingston upon Hull (2006)
14. Chapelle, O., Metzler, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: *Proceedings of the 18th International Conference on Information and Knowledge Management (CIKM 2009)*, pp. 621–630. ACM Press (2009)
15. Croft, W.B., Metzler, D., Strohman, T.: *Search Engines: Information Retrieval in Practice*. Addison-Wesley, Reading (2009)
16. Di Nunzio, G.M., Leveling, J., Mandl, T.: LogCLEF 2011 multilingual log file analysis: language identification, query classification, and success of a query. In: *CLEF 2011 Labs and Workshop, Notebook Papers* (2011)
17. Di Nunzio, G.M., Leveling, J., Mandl, T.: Multilingual log analysis: LogCLEF. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) *ECIR 2011*. LNCS, vol. 6611, pp. 675–678. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20161-5_68
18. Ferrante, M., Ferro, N., Maistro, M.: Injecting user models and time into precision via Markov chains. In: *Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2014)*, pp. 597–606. ACM Press (2014)

19. Ferro, N., Lucchese, C., Maistro, M., Perego, R.: On including the user dynamic in learning to rank. In: Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press (2017)
20. Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A.: User profiles for personalized information access. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web. LNCS, vol. 4321, pp. 54–89. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-72079-9_2
21. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005), pp. 154–161. ACM Press (2005)
22. Joachims, T., Swaminathan, A., Schnabel, T.: Unbiased learning-to-rank with biased feedback. In: de Rijke, M., Shokouhi, M., Tomkins, A., Zhang, M. (eds.) Proceedings of the 10th ACM International Conference on Web Searching and Data Mining (WSDM 2017), pp. 781–789. ACM Press (2017)
23. Keenoy, K., Levene, M.: Personalisation of web search. In: Mobasher, B., Anand, S.S. (eds.) ITWP 2003. LNCS (LNAI), vol. 3169, pp. 201–228. Springer, Heidelberg (2005). https://doi.org/10.1007/11577935_11
24. Koidl, K., Conlan, O., Wade, V.: Cross-site personalization: assisting users in addressing information needs that span independently hosted websites. In: Proceedings of the 25th ACM Conference on Hypertext and Social Media, pp. 66–76. ACM (2014)
25. Leveling, J., Di Nunzio, G.M., Mandl, T.: LogCLEF: enabling research on multilingual log files. In: Proceedings of the First Workshop on Personalised Multilingual Hypertext Retrieval, PMHR 2011, pp. 55–56. ACM Press (2011)
26. Lupu, M., Hanbury, A.: Patent retrieval. *Found. Trends Inf. Retr. (FnTIR)* **7**(1), 1–97 (2013)
27. Mandl, T., Agosti, M., Di Nunzio, G.M., Yeh, A., Mani, I., Doran, C., Schulz, J.M.: LogCLEF 2009: the CLEF 2009 multilingual logfile analysis track overview. In: Peters, C., Di Nunzio, G.M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G. (eds.) CLEF 2009. LNCS, vol. 6241, pp. 508–517. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15754-7_62
28. Mandl, T., Di Nunzio, G.M., Schulz, J.M.: LogCLEF 2010: the CLEF 2010 multilingual logfile analysis track overview. In: CLEF 2010 LABs and Workshops, Notebook Papers (2010)
29. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst. (TOIS)* **27**(1), 2:1–2:27 (2008)
30. Müller, H.: Medical (visual) information retrieval. In: Agosti, M., Ferro, N., Forner, P., Müller, H., Santucci, G. (eds.) PROMISE 2012. LNCS, vol. 7757, pp. 155–166. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36415-0_10
31. Spina, D., Maistro, M., Ren, Y., Sadeghi, S., Wong, W., Baldwin, T., Cavedon, L., Moffat, A., Sanderson, M., Scholer, F., Zobel, J.: Understanding user behavior in job and talent search: an initial investigation. In: Proceedings of the 2017 SIGIR Workshop on eCommerce (eCom 2017). CEUR-WS.org (2017)
32. Sweetnam, M., Siochru, M., Agosti, M., Manfioletti, M., Orio, N., Ponchia, C.: Stereotype or spectrum: designing for a user continuum. In: the Proceedings of the First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage, ENRICH (2013)

33. Sweetnam, M.S., Agosti, M., Orio, N., Ponchia, C., Steiner, C.M., Hillemann, E.-C., Ó Siochrú, M., Lawless, S.: User needs for enhanced engagement with cultural heritage collections. In: Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F. (eds.) TPDL 2012. LNCS, vol. 7489, pp. 64–75. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33290-6_8
34. Wu, Q., Burges, C.J.C., Svore, K.M., Gao, J.: Adapting boosting for information retrieval measures. *Inf. Retr.* **13**(3), 254–270 (2010)

Multimedia Content Analysis

An Abstract Argumentation-Based Approach to Automatic Extractive Text Summarization

Stefano Ferilli^(✉) and Andrea Paziienza

Dipartimento di Informatica, Università di Bari, Bari, Italy
{stefano.ferilli, andrea.paziienza}@uniba.it

Abstract. Sentence-based extractive summarization aims at automatically generating shorter versions of texts by extracting from them the minimal set of sentences that are necessary and sufficient to cover their content. Providing effective solutions to this task would allow the users of Digital Libraries to save time in selecting documents that may be appropriate for satisfying their information needs or for supporting their decision-making tasks. This paper proposes an approach, based on abstract argumentation, to select the sentences in a text that are to be included in its summary. The proposed approach obtained interesting experimental results on the English subset of the benchmark MultiLing 2015 dataset.

Keywords: Text summarization · Digital libraries
Abstract argumentation

1 Introduction

Text summarization aims at automatically creating a shorter version of (a set of) text document(s). Abstractive techniques [1] produce summaries that may contain sentences not present in the input document(s). Extractive techniques [14] select a subset of sentences from the input document(s). An accurate extractive summarization method must optimize two important properties [22]: *coverage*, expressing how much the method is able to cover a sufficient amount of topics from the original text, and *diversity*, which refers to the capability of the method of generating non-redundant information in the summary. Graph-based methods for automatic text summarization have provided encouraging results. Nodes in the graph are sentences in the input document(s), and weighted edges are placed whenever a node/sentence refers to another. Weights are used to generate the scores of sentences.

Argumentation is an inferential strategy that aims at selecting reliable items in a set of conflicting claims (the *arguments*). It has been a very active topic in Artificial Intelligence since more than two decades now. Specifically, the Abstract Argumentation Frameworks [6] work on graphs in which nodes represent arguments, and edges represent attack (or, sometimes, support) relationships among arguments. Several non-monotonic reasoning approaches have been defined, that

allow to understand which subsets of arguments in the graph are mutually compatible, based on the fact that they are able to defend each other from attacks of other disputing arguments. Some of these approaches may handle weighted graphs/edges.

This paper proposes an approach to extractive text summarization based on abstract argumentation. Attacks represent the fact that two sentences cannot be both included in the summary. *Vice versa*, supports represent the fact that two sentences should be both included in the summary. The set of ‘consistent’ arguments/sentences computed by argumentation should represent a suitable summary. Attacks and supports are set based on the degree of similarity between two sentences: indeed, different sentences are likely to cover a larger portion of the original text, while similar sentences are likely to bear much redundancy. In this perspective, the weight on the graph edges (i.e., the kind and strength of the relationship between sentences/arguments) might be determined using some similarity measure. In a nutshell, the underlying idea is to place an attack relation between pairs of sentences whose similarity is high, in order to enforce the diversity property. *Vice versa*, a support relation is introduced between pairs of sentences with low similarity, in order to enforce the coverage property.

This paper is organized as follows. The next two sections lay out the background of our research. Section 4 introduces our proposals, and Sect. 5 evaluates its performance. Finally, Sect. 6 concludes the paper.

2 Related Work on Text Summarization

Extractive Text Summarization methods are usually performed in three steps [18]: (i) creation of an intermediate representation of the input which captures only the key aspects of the text (by dividing the text into paragraphs, sentences, and tokens); (ii) scoring of the sentences based on that representation; and (iii) generation of a summary consisting of several sentences, selected by appropriate combination of the scores computed in the previous step.

The score of sentences should be computed using a measure that is able to express how significant they are to the understanding of the text as a whole. For instance, [10] proposed to score sentences using a new measure that expresses their similarity. Its computation encompasses three linguistic layers: (i) the lexical layer, which includes lexical analysis, stopwords removal and stemming; (ii) the syntactic layer, which performs syntactic analysis; and (iii) the semantic layer, that mainly describes the annotations that play a semantic role. These three layers handle the two major problems in measuring sentence similarity, i.e., the meaning and word order problems, in order to automatically combine different levels of information in the sentence while assessing similarity.

In this setting, many strategies have been proposed in the literature to determine which sentences in a given text can be considered as representative of its content. *Word scoring* approaches [12] assigns scores to the most important words. On the other hand, *sentence scoring* approaches determine the features of sentences by detecting and by leveraging the presence of cue-phrases [21] and

numerical data [9]. Finally, *graph scoring* analyzes the relationships between the sentences that make up the text. The TextRank algorithm [16] extracts important keywords from a text document, where the weight expressing the importance of a word within the entire document is determined using an unweighted graph-based model.

Some efforts spent in combining the various approaches prove that hybrid approaches may lead to better results. Indeed, [9] shows that combining scoring techniques leads to an improvement in the performance of both single- and multi-document summarization tasks, as measured by the traditional metrics used in this setting (ROUGE scores —see next). In general, the same techniques used in single document summarization systems are applicable to multi-document ones.

Finally, in order to form a paragraph length summary, one approach is based on *Maximal Marginal Relevance* [2], in which the best combination of important sentences is selected.

3 Abstract Argumentation

Since the approach to extractive text summarization that we will propose in this paper is based on argumentative reasoning, we first recall here the basics of this inference strategy. As said, argumentation is an inferential strategy that aims at selecting reliable items in a set of conflicting claims (the *arguments*). In particular, we will consider Abstract Argumentation, which neglects the actual content and inner structure of arguments, to focus just on their external relationships of attack (or, sometimes, support). These relationships may be expressed by a graph in which nodes represent arguments, and edges represent attack or support relationships among arguments.

One of the most influential computational models of arguments proposed in this setting is represented by Dung’s Argumentation Frameworks [6] (AFs for short), defined as follows.

Definition 1. *An Argumentation Framework (AF) is a pair $F = \langle \mathcal{A}, \mathcal{R} \rangle$, where \mathcal{A} is a finite set of arguments and $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$. The relation $a\mathcal{R}b$ means that a attacks b .*

There are a few central concepts when evaluating the justification of an argument:

Definition 2. *Let $F = \langle \mathcal{A}, \mathcal{R} \rangle$ be an AF and $S \subseteq \mathcal{A}$. Then:*

- S is conflict-free if $\nexists a, b \in S$ s.t. $a\mathcal{R}b$;
- $a \in \mathcal{A}$ is defended by S if $\forall b \in \mathcal{A}: b\mathcal{R}a \Rightarrow \exists c \in S$ s.t. $c\mathcal{R}b$;
- $f_F: 2^{\mathcal{A}} \mapsto 2^{\mathcal{A}}$ s.t. $f_F(S) = \{a \mid a \text{ is defended by } S\}$ is called the characteristic function of F .
- S is admissible if S is conflict-free and S is defended by itself, i.e. $\forall a \in S, \forall b \in \mathcal{A}: b\mathcal{R}a \Rightarrow \exists c \in S$ s.t. $c\mathcal{R}b$.

The conditions for arguments acceptance are defined by different semantics. Semantics produce acceptable subsets of the arguments, called *extensions*, that correspond to various positions one may take based on the available arguments. Standard acceptability semantics characterize admissible sets of arguments:

Definition 3. Let $F = \langle \mathcal{A}, \mathcal{R} \rangle$ be an AF and $S \subseteq \mathcal{A}$ be an admissible set. Then, S is a:

- complete extension iff $S = f_F(S)$;
- grounded extension iff S is the \subseteq -minimal complete extension.
- preferred extension iff S is a \subseteq -maximal complete extension.
- stable extension iff $\forall a \in \mathcal{A}, a \notin S, \exists b \in S$ s.t. bRa .

The justification state of an argument can be conceived in terms of its extension membership. A basic classification encompasses only two possible states for an argument, namely justified or not justified. In this respect, two alternative types of justification, i.e. skeptical or credulous, can be considered.

Definition 4 (Justification State). Let $F = \langle \mathcal{A}, \mathcal{R} \rangle$ be an AF, and $\mathcal{E}_\sigma(F) = \{S \subseteq \mathcal{A} \mid \sigma(S)\}$ be the set of extensions for a given semantics σ ($\sigma \in \{\text{complete, grounded, preferred, stable}\}$). Then, an argument $a \in \mathcal{A}$ is:

- skeptically justified iff $\forall E \in \mathcal{E}_\sigma(F) : a \in E$;
- credulously justified iff $\exists E \in \mathcal{E}_\sigma(F) : a \in E$.

Many strategies can be found in the literature for the identification of the successful arguments in an argumentation dispute [3, 7, 19].

A Bipolar AF (BAF) [3] is an extension of Dung's AF in which two kinds of interactions between arguments are possible: attack and support. These two relations are independent and lead to a bipolar representation of the interaction between arguments. A BAF can be represented by a directed graph in which two kinds of edges are used, in order to differentiate between the two relations.

Definition 5. A BAF is a triplet $B = \langle \mathcal{A}, \mathcal{R}_{att}, \mathcal{R}_{sup} \rangle$, where \mathcal{A} is a set of arguments, \mathcal{R}_{att} is a binary relation on \mathcal{A} called attack relation and \mathcal{R}_{sup} is another binary relation on \mathcal{A} called support relation. For two arguments a and b , $a\mathcal{R}_{att}b$ (resp., $a\mathcal{R}_{sup}b$) means that a attacks b (resp., a supports b).

In BAFs, new kinds of attack emerge from the interaction between the direct attacks and the supports: there is a *supported attack* for an argument b by an argument a iff there is a sequence of supports followed by one attack, while there is an *indirect attack* for an argument b by an argument a iff there is an attack followed by a sequence of supports. In particular, we say that a supports b if there is a sequence of direct supports from a to b . Taking into account sequences of supports and attacks leads to the following definitions applying to sets of arguments [3].

Definition 6. Let $B = \langle \mathcal{A}, \mathcal{R}_{att}, \mathcal{R}_{sup} \rangle$ be a BAF. A set $S \subseteq \mathcal{A}$ set-attacks an argument $b \in \mathcal{A}$, iff there exists a supported attack or an indirect attack for b from an element of S .

A set $S \subseteq \mathcal{A}$ set-supports an argument $b \in \mathcal{A}$, iff there exists a sequence $a_1 \mathcal{R}_{sup} \dots \mathcal{R}_{sup} a_n$, $n \geq 2$, such that $a_n = b$ and $a_1 \in S$.

A set $S \subseteq \mathcal{A}$ defends an argument $a \in \mathcal{A}$, iff for each argument $b \in \mathcal{A}$, if $\{b\}$ set-attacks a , then b is set-attacked by S .

In the following, we define the semantics for acceptability in BAFs.

Definition 7. Let $B = \langle \mathcal{A}, \mathcal{R}_{att}, \mathcal{R}_{sup} \rangle$ be a BAF and $S \subseteq \mathcal{A}$. Then, S is:

- conflict-free, iff $\nexists a, b \in S$ s.t. $\{a\}$ set-attacks b ;
- safe, iff $\nexists b \in \mathcal{A}$ s.t. S set-attacks b and either S set-supports b or $a \in S$;
- a d-admissible extension, iff S is conflict-free and $\forall a \in S$, a is defended by S ;
- an s-admissible extension, iff S is safe and $\forall a \in S$, a is defended by S ;
- a d-preferred (resp. s-preferred) extension is a \subseteq -maximal d-admissible (resp. s-admissible) subset of \mathcal{A} .

A weighted AF (WAF) [7] is another extension of Dung’s AF in which attacks between arguments are associated with a weight, indicating the relative strength of the attack. In this framework, some inconsistencies are tolerated in subsets S of arguments, provided that the sum of the weights of attacks between arguments of S does not exceed a given inconsistency budget $\beta \in \mathbb{R}_*^+$. The meaning is that attacks up to a total weight of β are neglected. Dung’s argument systems assume an inconsistency budget of 0, while, by relaxing this constraint, WAFs can achieve more solutions.

4 Argumentation-Based Text Summarization

Our summarization framework consists of the following phases:

Natural Language pre-processing. The document d to be summarized is progressively splits into sentences $\langle s_1, s_2, \dots, s_n \rangle$, then each sentence s_i is split into a sequence of tokens (words) $\langle s_{i,1}, s_{i,2}, \dots, s_{i,k} \rangle$, and finally each token undergoes lemmatization and stopword removal.

Weighted graph building. The similarity between each pair of sentences is computed and exploited to generate a weighted graph $G = (V, E, f_w)$ where nodes $V = \{s_1, s_2, \dots, s_n\}$ are the sentences in d , and edges $E \subseteq V \times V$ are weighted by the degree of similarity between the associated sentences as computed by the weighting function $f_w : E \rightarrow \mathbb{R}$.

Argumentation Framework building and evaluation. The resulting graph G is used to build an argumentation framework F_G expressing (possibly weighted) attacks and supports, on which computing a semantics that will determine which sentences are to be selected for inclusion in the summary.

Each phase can be implemented using different approaches and techniques. In the following we will focus on the last phase. Designing this phase requires to choose 3 components: (i) the Argumentation Framework setting; (ii) the graph transformation procedure that builds F_G starting from G ; and (iii) the semantics for computing the summary.

As to the first issue, in this preliminary investigation we focused on BAFs, as the simplest AF that allows to consider both attacks and supports between arguments. Concerning the second issue, to derive supports and attacks starting from the weighted edges in G , we defined a heuristic, inspired by the concept of *inconsistency budget* in weighted argumentation frameworks [7]. Intuitively, we want to consider as supports arcs connecting sentences which are dissimilar from each other (because, in some sense, they may bear disjoint information that is worth including in the summary in order to enforce the *coverage* property), and to set attacks between pairs of similar sentences (to model the fact that including both in the summary would not bring much additional information to the summary, violating the *diversity* property). So, we normalize to $[0, 1]$ the weights in G and define two thresholds in order to distinguish which edges in G are attacks or supports in F_G :

- the *attack threshold* $\alpha \in [0, 1]$ and
- the *support threshold* $\beta \in [0, 1]$,

with $\beta < \alpha$. Then, we generate $F_G = \langle \mathcal{A}, \mathcal{R}_{att}, \mathcal{R}_{sup} \rangle$ such that $\mathcal{R}_{sup} = \{e \in E \mid f_w(e) \leq \beta\}$, $\mathcal{R}_{att} = \{e \in E \mid f_w(e) \geq \alpha\}$, and $\mathcal{A} = \{v \in V \mid \exists u \in V : (v, u) \in \mathcal{R}_{sup} \cup \mathcal{R}_{att} \vee (u, v) \in \mathcal{R}_{sup} \cup \mathcal{R}_{att}\}$. So, we place attacks between very similar sentences, supports between very dissimilar ones, and leave an intermediate similarity range for which we do not set attacks nor supports. Finally, for the third issue, once the BAF is instantiated, we considered the following extension-based semantics (listed from the most credulous to the most skeptical ones) to evaluate the acceptability of arguments: *d-admissible*, *s-admissible*, *complete*, *d-preferred*, *s-preferred* and *stable*. What we expect from the arguments evaluation is that:

- *conflict-free* sets collect the largest subsets of arguments encompassing the main principle of argumentation solutions, i.e., the idea that winning arguments should not attack each other. This requirement would reward sentences that maximize the diversity property. However, for the coverage property, we require that a solution defends its element, too.
- *d-/s-admissible* sets collect a large number of arguments. In principle, these solutions may be appropriate for text summarization tasks. However, when the allowed length of the summary is constrained by an upper bound, admissible solutions may include too many arguments, thus yielding a summary which is ideally good but exceeds the allowed length.
- *complete* extensions collect sets of arguments which can defend all and only their elements from external attacks. They are still admissible and will achieve at most as many solutions as the admissible ones. However, complete semantics may include sets of arguments that are too small.

- *d/s-preferred* extensions collect sets of arguments maximally-included in *d/s*-admissible sets. Therefore, preferred semantics should in principle behave better than admissible and complete ones, both in terms of quality of the summary and in terms of summary length.
- *stable* extensions collect sets of arguments that are able to attack all the remaining arguments not included in the set. In terms of summary requirements, they contain the most dissimilar sentences. Due to their strong requirements, stable semantics might not achieve any solution at all.

5 Evaluation

The effectiveness of the proposed approach was evaluated on the *single-document text summarization* task of the English dataset of the *MultiLing 2015* challenge [11]. This allowed us to have both the ground truth and the state-of-the-art results, published after the competition, available for testing and comparison purposes. On average, the input texts were made up of about 25542 characters, while the ground truths were made up of about 1857 characters (i.e., about 7% of the source texts) on average. Following the experimental protocol defined for the challenge, two variants of the ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) [13] measure were used to quantitatively evaluate the generated summaries: *ROUGE-1* and *ROUGE-2*, where *ROUGE-N* is the *N*-gram recall between a candidate summary and the reference summary. So, *ROUGE-1* and *ROUGE-2* consider the ratio of co-occurring unigrams (i.e., single words) and bi-grams (i.e., pairs of adjacent words), respectively, in a candidate summary over the reference summaries.

In this respect, it is important to point out that the ground truth in this dataset is obtained by humans using an abstractive approach. This deserves some discussion. First, the comparison under these conditions is partly unfair, because there is no exact match between the sentences in the input text and those in the summary. Indeed, since the summary authors were free to merge and restructure in one sentence several parts of the input texts, possibly belonging to many different sentences, the chances of obtaining the same results are significantly affected. Even worse, when inspecting the summaries, it turned out that they may include words that were not present at all in the input texts (e.g., full spelling of the acronyms). This means that, using an extractive summarization approach, it might be impossible to match exactly the ground truth summary, even considering the whole input text. Last but not least, condensing the original content into just 7% opens significant possibilities that the authors of the summary have taken many subjective decisions about what to include in, and what to filter out from, the ground truth, leaving the possibility that other summaries might be as good as theirs, but quite different from theirs.

In the following, we compared our method with the following baselines, taken from the published results of the *MultiLing 2015* competition, whose performance is reported in Table 1:

Table 1. Experimental evaluation results for the MultiLing 2015 dataset

	ROUGE-1	ROUGE-2
WORST	37.17%	9.93%
BEST	50.38%	15.10%
ORACLE	61.91%	22.42%

WORST the worst-performing approach of the challenge;

BEST the best-performing approach of the challenge;

ORACLE an upper bound on the extractive text summarization performance: it uses a covering algorithm [5] that selects sentences from the original text covering the words in the summary disregarding the length limit.

These approaches are set so as to return summaries having the exact length in characters as the ground truth. This is questionable as well, since by truncating a candidate summary to a pre-defined number of characters spoils the very aims and motivations of summarization, which is returning a shorter version of the text that still conveys most of the original content and is human-understandable.

On the other hand, as already pointed out, argumentation semantics return subsets (‘extensions’) of arguments (sentences) that are mutually consistent (‘justified’). This means that, using our argumentation-based approach, (i) we have no control on the number of sentences that are selected to make up the summary, except for choosing different semantics that tend to return larger or smaller extensions; and (2) the control is at the level of sentences, whereas in the challenge length comparison to the ground truth is made in terms of characters. For semantics that returned several extensions, in the spirit of summarization, the shortest one was adopted.

Concerning the first two phases of the approach, we adopted the same solutions as in [8], that we will quickly recall in the following for the sake of illustration. The Natural Language pre-processing phase was mainly based on the *Stanford CoreNLP* toolkit [15], including the dependency parser [4] to extract additional information about word dependency. The *Simplified Lesk* algorithm [23] was also used for word sense disambiguation based on *Wikipedia* or *WordNet* [17], and word embeddings were computed. As regards graph building, the weight between two sentences is computed based on the similarity of their building tokens computed using a combination of three different similarity functions: a *syntactic similarity* based on the *Jaccard Index* applied to syntactic dependencies; a *semantic similarity* based on the function proposed in [20] for taxonomic information based on the synsets in *WordNet*, and an *embedding similarity* based on *cosine similarity* between the word embeddings.

As regards the argumentation phase, we built several argumentation frameworks by setting different values for thresholds α and β . Specifically, we considered values for the support threshold β ranging into [0.1, 0.8], and values for the attack threshold α ranging into [0.15, 0.9], using a 0.05 step for both and ensuring that $\beta < \alpha$ in each setting. All valid threshold combinations resulted

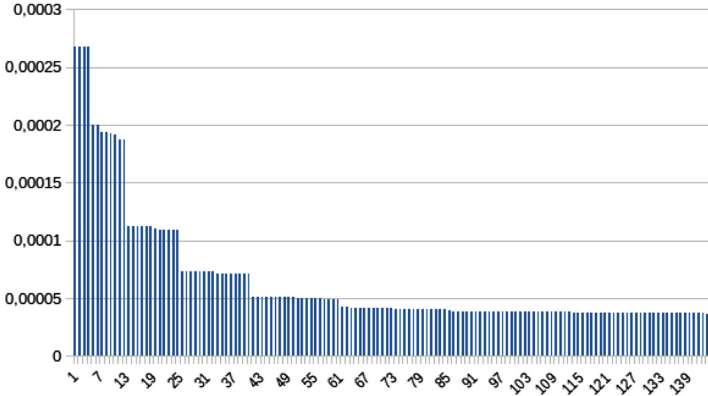


Fig. 1. Quality results for the different argumentation settings on the MultiLing 2015 dataset

in several hundreds of argumentation frameworks for each summarization task, on each of which we computed all the extension-based semantics selected in the previous section: d-/s-admissible, d-/s-preferred, stable and complete. Occasionally, we also computed conflict-free sets, that might provide the most suitable trade-off between acceptability membership conditions and justification state of arguments.

To have an immediate idea of the balance between the summarization performance and the length of a summary s , we defined a compound indicator:

$$\text{Quality}(s) = \text{ROUGE-1}(s)/\text{length}(s)$$

where the length of the summary, placed at the denominator, penalizes the quality of the solution s . Figure 1 graphically summarizes the results: each item on the x axis corresponds to a summarization task, for which the y axis reports the average Quality value obtained on all settings run for that task. Tasks are ordered on the x axis by decreasing average Quality. For the sake of clarity, the graph considers only 144 (σ, α, β) settings, that were selected as representative of the whole behavior. Specifically, $\sigma \in \{\text{s-admissible, d-admissible, stable, complete}\}$ is the semantics, and $\beta \in [0.1, 0.8]$ and $\alpha \in [0.2, 0.9]$, using a 0.1 step, are the thresholds in which the constraint $\beta < \alpha$ is satisfied for the semantics σ .

While in most of the graphic the decay in performance is smooth, it is also apparent that 4 ‘steps’ naturally emerged, associated to sudden drops in quality, as if a phase transition occurred. So, we leveraged these steps to select the most relevant settings to investigate in more depth. Each step corresponds to two different semantics that returned exactly the same results, as reported in Table 2. Note that summaries associated to the first step are shorter than the ground truth (1352 characters against 1857), then in the second step the length of the summaries jumps from 1352 to 4365 characters. So, there was no summary whose length was close to that of the ground truth (1857 characters). The ROUGE-1

Table 2. Experimental evaluation results for our approach. Average size of full texts is 25542 characters, average size of the ground truths is 1857 (7% of the full texts)

Step	Semantics	α	β	Length (%)	Quality	Rouge-1		Rouge-2	
						Recall	Precision	Recall	Precision
1	s-admissible d-admissible	0.1	0.3	1279 (5%)	2.00E-04	25.57%	41.97%		
2	s-admissible d-admissible	0.1	0.4	4365 (17%)	1.13E-04	49.28%	30.32%	15.49%	7.22%
3	stable complete	0.1	0.5	8544 (33%)	7.07E-05	60.44%	24.44%	23.98%	7.43%
4	s-admissible d-admissible	0.1	0.5	9826 (38%)	7.33E-05	72.09%	26.65%	27.26%	6.16%

results are comparable to the state-of-the-art at step 2, comparable to Oracle at step 3, and much ($>10\%$) better than Oracle at step 4, but using the 17%, 33% and 38% of the input texts, respectively (compared to 7% of the ground truth). In other words, the argumentation approach requires more than twice as many characters as the ground truth to obtain the same ROUGE-1 results as the state-of-the-art, and 1/3 of the whole text to reach Oracle. By allowing it to use a little more text, but however less than 2/5 of the input text, it is able to catch nearly 3/4 of the content. Considering ROUGE-2, the same comments as above still hold, but in this case the recall value is slightly larger than the reference systems. Also the results at step 1 are interesting: even if the length of the summary is less than that of the ground truth, recall is not so bad, and precision is quite high. ROUGE-2 was not computed for the first step, due to the summary being very short.

These results suggest that the proposed Argumentation-based approach is sensible and effective in returning relevant summaries, and is competitive in performance, albeit paying in summary length. Confirming our hypothesis, s-preferred and d-preferred semantics provide relevant results. However, also the stable and complete semantics may yield interesting results that somehow represent a trade-off corresponding to the performance of ORACLE.

Given the considerations about possible unfairness of the ground truth construction, we wanted to carry out also a qualitative evaluation of our summaries, by asking human beings to read them and provide their sensations. Very interestingly, they reported that the proposed summaries have little redundancy, yet provide a sensible account of the original document, also ensuring smooth discourse flow, even if they were obtained by filtering out sentences that, since present in the original text, presumably included relevant parts as regards the content and/or the flow of discourse.

6 Conclusion

The ever-increasing number of text documents that are present in digital libraries makes it impossible for humans to read and understand them in order to assess their relevance and/or grasp the content they express. Automatic text summarization is a possible solution, aimed at using computers to extract automatically summaries of the input text(s) that preserve their fundamental meaning. Thus, providing effective solutions to this task would bring enormous benefit to the library users. This paper focused on extractive text summarization, aimed at selecting subsets of sentences taken from the original documents that are necessary and sufficient to cover their content. Specifically, it proposed a framework whose core step is carried out using abstract argumentation, based on a similarity assessment between pairs of sentences in the document.

Experimental results obtained on the English subset of the benchmark MultiLing 2015 dataset confirmed the viability and effectiveness of the proposed approach. Differently from other approaches in the literature, the argumentation-based approach autonomously determines the number of sentences to be included in the summary, which is typically larger than required by the dataset's ground truth. However, the summaries are still significantly shorter than the original text, and reach very high performance. Being the approach general, we expect that similar results can be obtained on other languages, as well.

Future work will carry out further investigations on the possibility of improving the performance of the approach by exploring other argumentation frameworks (e.g., those that may handle weights on attacks and supports) and semantics. Also, further experiments on additional datasets and languages are planned.

Acknowledgments. This work was partially funded by the Italian PON 2007–2013 project PON02_00563_3489339 ‘Puglia@Service’.

References

1. Banerjee, S., Mitra, P., Sugiyama, K.: Multi-document abstractive summarization using ILP based multi-sentence compression. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence, IJCAI 2015, pp. 1208–1214 (2015)
2. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: ACM SIGIR, pp. 335–336. ACM (1998)
3. Cayrol, C., Lagasque-Schiex, M.C.: On the acceptability of arguments in bipolar argumentation frameworks. In: Godo, L. (ed.) ECSQARU 2005. LNCS (LNAI), vol. 3571, pp. 378–389. Springer, Heidelberg (2005). https://doi.org/10.1007/11518655_33
4. Chen, D., Manning, C.D.: A fast and accurate dependency parser using neural networks. In: EMNLP, pp. 740–750 (2014)
5. Davis, S.T., et al.: OCCAMS-an optimal combinatorial covering algorithm for multi-document summarization. In: ICDMW, pp. 454–463. IEEE (2012)

6. Dung, P.M.: On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n-person games. *Artif. Intell.* **77**(2), 321–357 (1995)
7. Dunne, P.E., et al.: Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artif. Intell.* **175**(2), 457–486 (2011)
8. Ferilli, S., Paziienza, A., Angelastro, S., Suglia, A.: A similarity-based abstract argumentation approach to extractive text summarization. In: Esposito, F., Basili, R., Ferilli, S., Lisi, F. (eds.) *AI*IA 2017*. LNCS, vol. 10640, pp. 87–100. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70169-1_7
9. Ferreira, R., et al.: Assessing sentence scoring techniques for extractive text summarization. *Expert Syst. Appl.* **40**(14), 5755–5764 (2013)
10. Ferreira, R., et al.: A new sentence similarity assessment measure based on a three-layer sentence representation. In: *DocEng*, pp. 25–34. ACM (2014)
11. Giannakopoulos, G., et al.: Multiling 2015: multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In: *SIGDIAL*, pp. 270–274 (2015)
12. Gupta, P., et al.: Summarizing text by ranking text units according to shallow linguistic features. In: *ICACT*, pp. 1620–1625. IEEE (2011)
13. Lin, C.: Rouge: A package for automatic evaluation of summaries. In: *ACL 2004 Workshop*, vol. 8 (2004)
14. Lloret, E., Palomar, M.: Text summarisation in progress: a literature review. *Artif. Intell. Rev.* **37**(1), 1–41 (2012)
15. Manning, C.D., et al.: The stanford CoreNLP natural language processing toolkit. In: *ACL (System Demonstrations)*, pp. 55–60 (2014)
16. Mihalcea, R., Tarau, P.: Texttrank: Bringing order into texts. *Association for Computational Linguistics* (2004)
17. Miller, G.: Wordnet: a lexical database for english. *Commun. ACM* **38**(11), 39–41 (1995)
18. Nenkova, A., McKeown, K.: A survey of text summarization techniques. In: Aggarwal, C., Zhai, C. (eds.) *Mining Text Data*, pp. 43–76. Springer, Heidelberg (2012). https://doi.org/10.1007/978-1-4614-3223-4_3
19. Paziienza, A., Esposito, F., Ferilli, S.: An authority degree-based evaluation strategy for abstract argumentation frameworks. In: *Proceedings of the 30th Italian Conference on Computational Logic*, pp. 181–196 (2015)
20. Rotella, F., Leuzzi, F., Ferilli, S.: Learning and exploiting concept networks with conNeKTion. *Appl. Intell.* **42**(1), 87–111 (2015)
21. Shardan, R., Kulkarni, U.: Implementation and evaluation of evolutionary connectionist approaches to automated text summarization. *J. Comput. Sci.* **6**, 1366–1376 (2010)
22. Umam, K., et al.: Coverage, diversity, and coherence optimization for multi-document summarization. *Jurnal Ilmu Komputer dan Informasi* **8**(1), 1–10 (2015)
23. Vasilescu, F., Langlais, P., Lapalme, G.: Evaluating variants of the lesk approach for disambiguating words. In: *LREC* (2004)

On Frequency-Based Approaches to Learning Stopwords and the Reliability of Existing Resources — A Study on Italian Language

Stefano Ferilli^(✉) and Floriana Esposito

University of Bari, Bari, Italy
{stefano.ferilli,floriana.esposito}@uniba.it

Abstract. Natural Language Processing techniques are of utmost importance for the proper management of Digital Libraries. These techniques are based on language-specific linguistic resources, that might be unavailable for many languages. Since manually building them is costly, time-consuming and error-prone, it would be desirable to learn these resources automatically from sample texts, without any prior knowledge about the language under consideration. In this paper we focus on stopwords, i.e., terms that can be ignored in order to understand the topic and content of a document. We propose an experimental study on the frequency behavior of stopwords, aimed at providing useful information for the development of automatic techniques for the compilation of stopword lists from a corpus of documents. The reliability and/or deficiencies of the stopwords obtained from the experiments is evaluated by comparison to existing linguistic resources. While the study is conducted on texts in Italian, we are confident that the same approach and experimental results may apply to other languages as well.

Keywords: Natural Language Processing · Linguistic resources
Stopwords · Keyword extraction

1 Introduction

In spite of the ever-growing spread of multimedia content in digital format, text is still the main channel by which information is represented, exploited and exchanged by humans. Accordingly, the overwhelming majority of content in Digital Libraries (DLs for short) is still in the form of text. In fact, often even non-textual documents are indexed and annotated based on a textual description of their content, also due to the complexity of extracting explicit information from them. It is likely that this landscape will never change significantly, because natural language is the tool that humans have developed and refined through the millenniums to express their thoughts and notions to other people. In turn, the availability of a huge number of texts in DLs and other kinds of digital repositories naturally causes the so-called *information overloading* problem and raises the issue of how to efficiently and effectively manage them, especially

for retrieval and consultation purposes. Indeed, manual management is clearly beyond human capabilities.

In order to solve this problem, research in Computer Science has started several research areas, the most fundamental of which (because all others rely on its results) is Natural Language Processing (NLP), aimed at developing advanced tools for understanding the components, structure and meaning of texts, and to properly organize this information so as to help human users in satisfying their information needs. Typical NLP tasks, ranging from the morphological level, through the lexical one, up to the syntactic and, for some limited applications, the semantic one, are the following:

- Language Identification** aims at discovering the language in which the text is written;
- Stopword Removal** removes the terms that are not informative about the specific text content;
- Normalization** standardizes to a single form inflected forms of words;
- Part-of-Speech Tagging** associates to each term its grammatical function;
- Parsing** returns the syntactic structure of sentences;
- Understanding** aims capturing some kind of semantic information underlying the text.

Each of these steps is typically carried out using suitable linguistic resources. Language Identification often exploits n -gram distribution, Stopword Removal exploits lists of frequent terms, Normalization exploits lists of suffixes (e.g., [18]), Part-of-Speech Tagging exploits suffixes and/or grammatical rules (e.g., [2]), Parsing uses grammars, Word Sense Disambiguation uses conceptual taxonomies or ontologies.

Since languages are very different from each other, these resources are necessarily language-specific. Unfortunately, developing these resources poses several issues. Each language has its own peculiarities, and hence the resources developed for a language are useless for the others. Manually designing and developing these resources by experts is a costly, time-consuming and error-prone activity. Once the resources are available, it is very hard to maintain and update them, or to tailor them to specific domains, or to fix possible errors. Most works in the literature are concerned with English, probably due to its having a structure which is easier than other languages and to its importance as the standard information interchange language worldwide. Little exists for a few other important languages, and almost nothing for the vast majority of minor languages. As a result, automatic high-level processing techniques cannot be applied to documents in these languages, leading to the risk that entire cultures might be lost.

A possible solution is trying to learn the resources and other useful linguistic information (semi-)automatically from texts in the various languages. Some attempts can be found in the literature for Language Identification (in the form of statistics on the distribution of n -grams across the various languages [1, 15, 16]), Part-of-Speech Tagging (e.g., by learning tagging rules [3, 4]), Parsing (with the research stream concerning grammar inference [6]) and Understanding (with initial attempts to learn concept taxonomies or graphs, or even ontologies, but often

based on existing taxonomies/graphs and/or semi-automatically [5, 11, 13, 14, 17, 21, 22]). Our contribution in this landscape was *BLA-BLA* (an acronym for ‘Broad-spectrum Language Analysis-Based Learning Application’), a tool that currently includes several techniques that allow to learn in a fully automatic way linguistic resources for language identification [8], stopword removal and term normalization [7, 9] and concept extraction [12, 19]. The learned resources may be used as returned by the system, and/or be taken as a basis for further manual refinements. Whenever more texts become available for the language, it is easy to run again the technique and obtain updated resources.

Stopwords, in particular, are terms in a language that appear so often and pervasively in the documents as to make them irrelevant to distinguish documents with respect to their content. From an information retrieval perspective, a stopword can be defined as a “word that has the same likelihood of occurring in those documents not relevant to a query as in those documents relevant to the query” [23]. So, by definition, stopwords can be safely ignored by NLP techniques that work at the lexical level. The removal task is simply carried out by look-up in a pre-determined list of words. The usual way for preparing such a list is including all *function words*, i.e. terms associated to invariant Parts-of-Speech of the language (usually articles, pronouns and prepositions), but this requires prior knowledge about the grammar of the language. Moreover, for domain-specific applications, also other terms that are insignificant in the particular context (e.g., the word ‘*computer*’ in a DL specialized in Computer Science) can be added to the list. For instance, [23] adopts this perspective, using a Vector Space Model to identify stopwords. However, the proposed technique applies Porter’s stemmer [18] prior to the stopword extraction step, which makes the approach language-dependent, and requires, again, the existence of tools/resources for that language. Two more purely frequency-based approaches are proposed in [10, 20]. However, the former still deals with English, and the latter specifically focuses on French. The former was tested on a corpus of broad literature including more than 1 million words, and the latter on two corpora made up of many small texts, but totaling more than 4 and more than 6 million words, respectively. Moreover, both manually adjusts the automatically determined list of stopwords. BLABLA aims at avoiding all these requirements and limitations: it uses just plain texts in a given language for learning, it requires very small corpora, it is fully automatic, it does not focus on a specific language.

In BLA-BLA, stopwords are currently identified as those terms that appear with a higher frequency than the other words in the training documents. The selection is based on a frequency threshold, that in the current prototype is simply set as the average frequency of all terms collected for the language, multiplied by a smoothing factor. So, in this paper, we focus on the automatic learning of stopword lists, with the aim of improving the technique embedded in BLABLA. More specifically, here we present a study of the frequency behavior of terms extracted from texts in a given language, depending on the type and amount of text, and on the mix of texts used for learning. Our study compares the experimental results with standard linguistic resources currently available, and discusses the critical issues

arising from such a comparison, both from the learning perspective and as regards the reliability of the existing resources. The learned lessons are then used to suggest how to possibly improve the approach of BLABLA to learn stopwords, and how to even extract keywords along with stopwords. Section 2 proposes a study of the behavior of frequent words in single texts or small corpora from the perspective of stopword learning. Then, Sect. 3 discusses the outcomes of the study, and Sect. 4 makes a proposal for keyword extraction. Lastly, Sect. 5 concludes the paper and outlines future work directions.

2 Experimental Study

BLA-BLA processes a set of input training documents in pure text, each of which is associated to the corresponding language. It assumes that each document belongs exactly to one language. This does not mean, of course, that it cannot include words or expressions from other languages, but these are to be considered as noise, and suitably handled by the learning approaches.

Concerning the lexical level, a pre-processing step is needed to extract from each document only *words*. In BLABLA, a word is formally defined by the following linear expression pattern:

$$\text{\textasciitilde}P\{W'\}^*WP\text{\textasciitilde}$$

where

- \textasciitilde is the blank symbol;
- $'$ is the apostrophe;
- $P = \{.,|,|;| :|?|!|''|\}^*$ is a (possibly empty) sequence of punctuation marks;
- $W = \{a|b|\dots|z\}^+$ is the word (hypothesizing a latin alphabet).

So, a word is a sequence of alphabetic characters only, delimited by blank spaces. Between the initial blank and the first character, and/or between the last character and the final blank, punctuation symbols are allowed (not considered as belonging to the word). The case of an apostrophe joining two words was considered as well.

Once the words only are extracted from a text (or a set of texts) T , they are collected in a multiset W . Then, for each word $w \in W$, its relative frequency is computed as the ratio of its number of occurrences over the overall number of word occurrences in the text(s): $f(w) = k/|W|$, where $|w|_W = k$ (i.e., w has k occurrences in T). Now, members of the *vocabulary* V (i.e., the set of unique words in W) are ranked by decreasing frequency. In a very simple (and simplistic) approach, the problem of determining the stopwords in T may be cast as the problem of cutting this list in such a way that all items above the cut point are considered as stopwords, and all items below the cut point are considered as relevant words. In turn, the cut point may be specified as a frequency threshold \bar{f} , such that all words $v \in V$ for which $f(v) \geq \bar{f}$ are considered as stopwords.

For our study, we focused on the Italian language, as an example of a language that has attracted some attention from the NLP community, albeit not as much

as English. So, existing stopword lists for this language may serve as a golden standard on which basing our study. It is also a language having a more complex structure than English, so that one may expect that, if good results are obtained on Italian, then good results might be obtained on most other languages, as well. We wanted to study the case in which only a small corpus is available for learning (say, involving just 10 texts). This should obviously stress the learning approach, because we may expect that, processing a large number of texts, the frequency of real stopwords will in the end become clearly predominant over the other words. We adopted this setting because for some languages (e.g., dialects) only very few written texts are available, because they live mostly in oral conversation.

So, we selected 10 texts from the Project Gutenberg¹ and Liber Liber² repositories, which make freely available for download many well-known texts from the literature of several languages. It should be noted that these texts are obtained by applying Optical Character Recognition to scanned images of paper books' pages, and so they contain spelling errors spread through the text, that introduce some noise. This is not necessarily bad, since it allows us to test our approaches also on noisy data, which are what one may expect to have in real-world settings.

Texts were selected so as to ensure a wide range of styles, and to support the study of frequency behavior when increasing the number of texts under different conditions. Specifically, we considered the following texts, where a letter in parentheses identifies the source from which the text was downloaded (G = Project Gutenberg, L = Liber Liber):

La Divina Commedia by Dante Alighieri (G), a poem written in the XIV century;

Codice Civile by the Italian Administration, a technical text of the XX century;

L'Esclusa by Luigi Pirandello (L), a novel written in the second half of the XIX century;

I Promessi Sposi by Alessandro Manzoni (L), a novel written in the first half of the XIX century;

Tutte le novelle by Giovanni Verga (L), a collection of stories written across the XIX and XX centuries;

Passeggiate per l'Italia (volumes 1–5) by Ferdinando Gregorovius (G), a description of travels made in the XIX century.

Table 1 reports some statistics about the length (in number of characters and of words) of the selected texts, and their linguistic variety (column 'Vocabulary' reporting the number of different words in each text). The number of characters and words is approximate (counted by a text editor), while the size of the vocabulary is exact (computed by the pre-processing step).

As the linguistic resource to be used as a golden standard, we chose the stopword list provided by Snowball³, a well-known tool exploited by many systems

¹ <https://www.gutenberg.org/>.

² <https://www.liberliber.it/>.

³ <http://snowball.tartarus.org/algorithms/italian/stop>.

Table 1. Statistics on the processed texts

#	Text	Chars	Words	Vocabulary
1	La Divina Commedia	561149	97714	12796
2	Codice Civile	1511666	228251	8659
3	L'Esclusa	337589	55846	8919
4	I Promessi Sposi	1307423	220174	19658
5	Tutte le novelle	1591823	264703	21641
6	Passeggiate per l'Italia 1	438868	71467	11995
7	Passeggiate per l'Italia 2	549884	86818	14710
8	Passeggiate per l'Italia 3	478110	75871	12721
9	Passeggiate per l'Italia 4	472272	75618	12183
10	Passeggiate per l'Italia 5	289006	46655	10470
11	Passeggiate per l'Italia	2228140	356429	30855

Table 2. Performance on the processed texts

Text(s) #	1	2	3	4	5	6	7	8	9	10	6-10	All	N-T
P@10	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
P@20	0.85	0.95	0.95	0.95	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
P@30	0.83	0.87	0.93	0.90	1.0	1.0	0.97	0.93	1.0	1.0	0.97	0.97	0.97
P@40	0.80	0.88	0.85	0.85	0.90	0.95	0.95	0.93	0.93	0.93	0.95	0.95	0.95
P@50	0.74	0.76	0.72	0.80	0.90	0.94	0.92	0.90	0.88	0.92	0.94	0.90	0.90
P@60	0.67	0.68	0.70	0.73	0.85	0.88	0.87	0.90	0.83	0.88	0.90	0.88	0.87
P@70	0.64	0.61	0.69	0.73	0.77	0.83	0.81	0.83	0.81	0.81	0.86	0.83	0.86
P@80	0.60	0.58	0.70	0.70	0.69	0.79	0.76	0.75	0.79	0.76	0.83	0.80	0.81
P@90	0.58	0.54	0.66	0.67	0.66	0.73	0.73	0.73	0.76	0.71	0.78	0.74	0.76
P@100	0.53	0.53	0.62	0.65	0.62	0.73	0.71	0.68	0.71	0.66	0.72	0.72	0.72
P = 1	11	5	14	14	30	33	27	21	33	30	28	27	25
R@100	0.19	0.19	0.22	0.23	0.22	0.26	0.25	0.24	0.25	0.24	0.26	0.26	0.26
P=R@279	0.32	0.28	0.35	0.38	0.38	0.36	0.36	0.36	0.34	0.36	0.37	0.40	0.41

currently in use. In its complete form, it consists of 279 stopwords. Table 2 reports the performance, in terms of precision (P), obtained on each single text ($\#$ as per Table 1) and on relevant aggregates of texts ('6-10' = the whole 'Passeggiate per l'Italia'; 'All' = the whole set of texts; 'N-T' = the subset of non-technical texts, excluding 'La Divina Commedia' and 'Codice Civile') when cutting the list at positions that are multiples of 10: so, $P@n$ means precision of the top n items in the ranking (i.e., the percentage of the first n items in the list that are also in the golden standard). These values may be considered as representative of the overall trend, albeit it is worth saying that the behavior of P is not monotonic, so some fluctuations are possible between the reported values. We considered only the first 100 positions both for the sake of readability, and because they represent a 'safety region' where most relevant stopwords should be included.

Table 2 also reports further information. $P = 1$ is the maximum position in the ranking at which 100% precision is preserved. $R@100$ is the recall value (R) associated to the last position in the ‘safety region’ (100), to compare it to $P@100$ and to have an idea of how much of the golden standard is still missing at that point in the list. As a reference, given that the golden standard stopword list includes 279 items, the maximum recall reachable @100 is $100/279 = 0.36$. Finally, $P = R@279$ reports the the performance at position 279. Note that, at this position, precision and recall take the same value ($P = R$).

3 Discussion

Table 2 shows that the length of the text is not strictly related to performance. The best performance is obtained on some volumes of ‘Passeggiate per l’Italia’, which is written in a kind of journalistic style. A slightly lower, but still quite high performance, is obtained on the stories of ‘Tutte le novelle’. The two novels come immediately after, followed by the two texts written using more particular styles, i.e., ‘Codice Civile’ (technical) and ‘La Divina Commedia’ (poetry). We may conclude that the writing style counts more than the number of words in the text, which makes sense but was partly unexpected. Specifically, colloquial styles are more useful for finding stopwords than technical ones. As expected, using many texts improves performance. While the improvement may not be outstanding compared to some single texts (e.g., 5 and 6), especially for the upper part of the ranking, a smoother decay in performance is clearly visible, as confirmed by the neat increase in performance @279.

The following is the list of errors, i.e., of words appearing in the top 100 items of each text but not in the golden standard, listed by decreasing frequency:

La Divina Commedia ch sì de d s quel me poi così m là quando già tanto son altro qual occhi ben disse sé lor qui ché or fa né com vidi n ogni elli pur però esser ciò giù altra tal prima ancor poco mondo te sù onde mai;

Codice Civile art può essere seguenti deve diritto cod contratto società caso civ beni disposizioni quando stato atto comma cosa parte secondo termine d possono salvo diritti codice legge titolo att devono altri azioni senza norme atti creditore fondo debitore terzo proc ogni valore parti luogo amministratori n persona;

L’Esclusa marta d s così occhi madre ora maria quasi no poi me quel sì via due casa signora egli dopo senza anna rocco alvignani ella marito mano ancora qua sotto ogni ah prima già disse giorno mani nulla;

I Promessi Sposi d quel s così disse poi renzo cosa de altro due qualche quando ora don senza ogni far lucia fatto parte tempo tanto bene gran qui ch altri casa fare dire uomo sempre già dopo;

Tutte le novelle d occhi quel quando senza altro poi ora fra due ella s casa tanto colle colla sotto ogni disse così cosa mani fatto prima egli capo dopo mano sempre tutta giorno dietro nulla quasi volta ancora né;

Passeggiate per l’Italia I d città roma ancora qui mare castello fra monti s due quando dopo ora tempo quasi così perchè campagna poi parte chiesa là strada prima ogni stato;

Passeggiate per l'Italia 2 roma d ebrei città chiesa impero tempo s due fra così quando sotto grande *ancora* ora storia tevere ogni parte stato già popolo egli quel essa dopo italia papa;

Passeggiate per l'Italia 3 roma d egli città italia così parte tempo *ancora* fra stato grande napoleone dopo s ravenna francia due papa essi solo già chiesa avignone ora romani quali storia senza quando garibaldi essere;

Passeggiate per l'Italia 4 d napoli città isola s due re mare sicilia quali tutte ogni così dopo fra popolo parte tutta *ancora* capri sotto senza palermo pure grande quasi quando siracusa quel;

Passeggiate per l'Italia 5 così d città s ora mare quando arte egli tempo vita perchè sempre già solo *ancora* sicilia intorno ciò due ogni casa tempio cuore allora essa dopo arrio popolo mentre euforione amore verso pompei;

Passeggiate per l'Italia d roma città così s due fra *ancora* tempo egli quando dopo ora parte ogni chiesa grande sotto mare quali italia stato già qui quel tutte solo senza;

Whole corpus d art s quel quando così può due poi senza altro essere cosa ogni ora ch parte tempo dopo prima stato occhi disse de tanto altri fatto sì;

Non-technical texts d quel s così quando due poi ora senza altro ogni dopo tempo cosa disse *ancora* città tanto egli casa fra prima sempre sotto fatto roma parte.

For the sake of subsequent discussion, we will consider as stopwords all words that do not have a definite meaning, indicating an object or an action, by themselves. According to this perspective, not only articles, pronouns, conjunctions and prepositions are stopwords, but also some adverbs and some verbs (e.g., modal verbs). Based on this definition, we underlined in the previous lists the words that we think are real errors. Words in italics are ambiguous, and can be considered as stopwords or not depending on how one interprets them. For instance, ‘stato’ may be a noun (‘state’), and thus it would not be a stopword, or the past participle of verb ‘to be’, and thus it would be a stopword. Similarly, ‘colla’ may mean ‘glue’ or it may be a contraction of ‘con la’; ‘colle’ may mean ‘hill’ or it may be a contraction of ‘con le’; ‘ancora’, depending on its accent, may mean ‘anchor’ or ‘still, again’; ‘ora’ may mean ‘hour’ or ‘now’; etc.

Since the above lists include a very large number of words that we would safely consider as stopwords, a question arises about the reliability and completeness of the state-of-the-art resources currently used for stopword removal (at least, for Italian). Actually, it is quite strange that some of these stopwords are not in the list used as the golden standard. Some examples: ‘essere’ is the infinitive form of verb ‘to be’, for which many inflected form are in the list; ‘fra’ is a very common alternate form of preposition ‘tra’, which is in the list; etc. More in general, many pronouns and generic adverbs are in these list, but not in the golden standard, even if it does include other similar pronouns or generic adverbs. If the stopwords in the above lists were added to the count of correct items, the results reported in Table 3 would be obtained, where ‘Loose’ considers the terms in italics as stopwords, and ‘strict’ considers them as non-stopwords (both the count of such terms, and the resulting precision @100, are reported).

Table 3. Performance on the processed texts: P@100

Text(s)	1	2	3	4	5	6	7	8	9	10	6–10	All	N-T
Original	0.53	0.53	0.62	0.65	0.62	0.73	0.71	0.68	0.71	0.66	0.72	0.72	0.72
Loose	4	30	14	10	7	10	13	15	12	14	8	5	7
P@100	0.96	0.70	0.86	0.90	0.93	0.90	0.87	0.85	0.88	0.86	0.92	0.95	0.93
Strict	0	1	2	1	4	3	3	3	1	2	3	2	2
P@100	0.96	0.69	0.84	0.89	0.89	0.87	0.84	0.82	0.87	0.84	0.89	0.93	0.91

A further insight may reveal other interesting details. From the perspective of texts:

- The poem includes many stopwords in truncated form (e.g., ‘d’ for ‘di’, ‘ch’ for ‘che’, etc.), which could be expected. Considering as correct these stopwords, the increase in performance would make this text the best one, instead of the worst.
- The technical text includes many specific words, which again could be expected. However, it does not include many stopwords, because they are seldom used in the specific domain. In facts, it still is the worst performing one, even after the corrections are applied.
- Nevertheless, including the technical text in the overall computation yields an improvements over the results obtained on non-technical texts only.
- After corrections, novels become the best-performing non-technical single texts: they reach the same performance as the ‘journalistic’ text(s) in the strict setting, and are even better than them in the loose setting.

From the perspective of terms/stopwords:

- Wrong terms in the lists associated to sets of texts are pushed towards the end of the list, confirming that larger corpora improve the quality of the results.
- Some terms appear in all lists, suggesting that they are actually stopwords that the golden standard failed to include (e.g., ‘d’, a truncation of preposition ‘di’).
- Some terms appear in the majority of lists (e.g., ‘quando’, ‘così’ appear in 9 texts; ‘dopo’, ‘due’, ‘ogni’ appear in 8 texts; ‘ora’, ‘ancora’ appear in 7 texts; ‘già’, ‘parte’, ‘quel’, ‘senza’ appear in 6 texts).
- Some terms in italics are in almost all lists, suggesting that they should be considered as stopwords (e.g., ‘ora’ and ‘ancora’ appear in 7 texts out of 10).

Finally, interesting considerations may be made also from the perspective of terms that are not stopwords. In facts, it is apparent that they might act as keywords for the corresponding texts: just by reading them one may infer that

- *La divina commedia* is a poem due to the presence of many truncated words;
- *Codice Civile* is about regulations and agreements among people;

- ‘I Promessi Sposi’ and ‘L’esclusa’ are novels, due to the presence of persons’ nouns (their main characters are clearly highlighted, indeed); in particular, L’Esclusa is about family relationships;
- *Passeggiate per l’Italia* is about geography/landscape, history/politics and art; more precisely, the first three volumes concern Rome, while the last two concern the Reign of the Two Sicilies, including Southern Italy and Sicily.

4 Proposal

Based on the above consideration, our proposal for extending BLABLA by improving its stopword extraction feature and adding a keyword extraction feature is the following. Given a set of texts, the frequency-based approach is used to extract candidate stopwords. If only one text is to be processed, it is likely that the resulting list will contain domain-specific terms, but they might be considered as domain-specific stopwords, according to the literature. If several texts are processed, a comparison of the stopwords extracted from the complete corpus to the stopwords extracted from the single texts may be used both to identify real stopwords and to extract keywords describing the specific content of the single texts. Applying this approach to the above lists would yield the following differences of the words found for the single texts with respect to those found for the whole corpus (‘All’):

La Divina Commedia altra ancor ben ché ciò com elli esser fa già già lor là
m mai me mondo n né ogne onde or per poco pur qual qui son s s tal te vidi;

Codice Civile amministratori att atti atto azioni beni caso civ cod codice
comma contratto creditore debitore deve devono diritti diritto disposizioni
fondo legge luogo n norme parti persona possono proc salvo secondo seguenti
società termine terzo titolo valore;

L’Esclusa ah alvignani ancora anna casa egli ella giorno già madre mani mano
maria marito marta me no nulla qua quasi rocco signora sotto via;

I Promessi Sposi bene casa dire don far fare già gran lucia qualche qui renzo
sempre uomo;

Tutte le novelle ancora capo casa colla colle dietro egli ella fra giorno mani
mano nulla né quasi sempre sotto tutta volta;

Passeggiate per l’Italia 1 ancora campagna castello chiesa città fra là mare
monti perchè quasi qui roma strada;

Passeggiate per l’Italia 2 ancora chiesa città ebrei egli essa fra già grande
impero italia papa popolo roma sotto storia tevere;

Passeggiate per l’Italia 3 ancora avignone chiesa città egli essi fra francia
garibaldi già grande italia napoleone papa quali ravenna roma romani solo
storia;

Passeggiate per l’Italia 4 ancora capri città fra grande isola mare napoli
palermo popolo pure quali quasi re sicilia siracusa sotto tutta tutte;

Passeggiate per l'Italia 5 allora amore ancora arrio arte casa città ci cuore
egli essa euforione già intorno mare mentre perch pompei popolo sempre sicilia
solo tempio verso vita;

Passeggiate per l'Italia ancora chiesa città egli fra già grande italia mare
quali qui roma solo sotto tutte.

We think that these results are sensible, especially considering the effort needed to obtain them. Indeed, differently from other techniques for stopword and keyword extraction proposed in the literature, for which heavy computations are required (e.g., to build version spaces based on TF*IDF-like schemes, or to compute statistics about co-occurrences of terms), our approach just requires a simple frequency count and a few set operations on lists of terms.

5 Conclusions and Future Work

Most content in Digital Libraries is still in the form of text, and this predominance will probably never be questioned. Except pure display of these documents, all other tasks are based on some kind of Natural Language Processing, that must be supported by suitable linguistic resources. Since these resources are clearly language-specific, they might be unavailable for several languages, and manually building them is costly, time-consuming and error-prone.

This paper studied the behavior of frequent words in single texts and (small) corpora and, based on the study, proposed a methodology to automatically learn a stopword list for a natural language starting from texts written in that language. The learned list may enable further high-level processing of documents in that language, and/or be taken as a basis for further manual refinements. The study suggested also that relevant keywords may be extracted from the texts with a little extension of the proposed approach. Preliminary experimental results show that the extracted stopwords and keywords are appropriate, and pointed out some deficiencies of standard resources available in the literature.

A future work issue is to define an approach to determine the threshold at which distinguishing stopwords from non-stopwords. Also, a study of the behavior on larger and more varied corpora should be carried out. Finally, an indirect evaluation of the quality of results through the evaluation of the performance of high-level NLP tasks based on the learned resources might be interesting.

References

1. Ahmed, B., Cha, S.-H., Tappert, C.: Language identification from text using n-gram based cumulative frequency addition. Proceedings of Student/Faculty Research Day, CSIS, Pace University, p. 12-1 (2004)
2. Brill, E.: A simple rule-based Part of Speech tagger. In: HLT 1991: Proceedings of the Workshop on Speech and Natural Language, pp. 112–116 (1992)
3. Brill, E.: Some advances in transformation-based Part of Speech tagging. In: Proceedings of the 12th National Conference on Artificial Intelligence (AAAI), vol. 1, pp. 722–727 (1994)

4. Brill, E.: Unsupervised learning of disambiguation rules for Part of Speech tagging. In: *Natural Language Processing Using Very Large Corpora Workshop*, pp. 1–13. Kluwer (1995)
5. Cimiano, P., Hotho, A., Staab, S.: Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Int. Res.* **24**(1), 305–339 (2005)
6. D’Ulizia, A., Ferri, F., Grifoni, P.: A survey of grammatical inference methods for natural language learning. *Artif. Intell. Rev.* **36**(1), 1–27 (2012)
7. Ferilli, S., Esposito, F., Grieco, D.: Automatic learning of linguistic resources for stopword removal and stemming from text. *Procedia Comput. Sci.* **38**, 116–123 (2014)
8. Ferilli, S., Esposito, F., Redavid, D.: Language identification as process prediction using woman. In: *Proceedings of the 12th Italian Research Conference on Digital Library Management Systems (IRCDL 2016)*, p. 12 (2016)
9. Ferilli, S., Grieco, D., Esposito, F.: Automatic learning of linguistic resources for stopword removal and stemming from text. In: Agosti, M., Ferro, N. (eds.) *Proceedings of the 10th Italian Research Conference on Digital Library Management Systems (IRCDL 2014)*, p. 12 (2014)
10. Fox, C.: A stop list for general text. *SIGIR Forum* **24**(1–2), 19–21 (1989)
11. Hensman, S.: Construction of conceptual graph representation of texts. In: *Proceedings of the Student Research Workshop at HLT-NAACL 2004, HLT-SRWS 2004*, pp. 49–54. Association for Computational Linguistics (2004)
12. Leuzzi, F., Ferilli, S., Rotella, F.: ConNeKTion: a tool for handling conceptual graphs automatically extracted from text. In: Catarci, T., Ferro, N., Poggi, A. (eds.) *IRCDL 2013. CCIS*, vol. 385, pp. 93–104. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-54347-0_11
13. Maedche, A., Staab, S.: Mining ontologies from text. In: *EKAW*, pp. 189–202 (2000)
14. Maedche, A., Staab, S.: The text-to-onto ontology learning environment. In: *ICCS-2000 - Eight International Conference on Conceptual Structures, Software Demonstration* (2000)
15. Martins, B., Silva, M.J.: Language identification in web pages. In: *Proceedings of the 2005 ACM Symposium on Applied Computing*, pp. 764–768. ACM (2005)
16. Nagarajan, T., Murthy, H.A.: Language identification using parallel syllable-like unit recognition. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, vol. 1, p. I-401. IEEE (2004)
17. Ogata, N.: A formal ontology discovery from web documents. In: Zhong, N., Yao, Y., Liu, J., Ohsuga, S. (eds.) *WI 2001. LNCS (LNAI)*, vol. 2198, pp. 514–519. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-45490-X_66
18. Porter, M.F.: An algorithm for suffix stripping. *Program* **14**(3), 130–137 (1980)
19. Rotella, F., Leuzzi, F., Ferilli, S.: Learning and exploiting concept networks with connexion. *Appl. Intell.* **42**, 87–111 (2015)
20. Savoy, J.: A stemming procedure and stopword list for general french corpora. *J. Assoc. Inf. Sci. Technol.* **50**, 944–952 (1999)
21. Shamsfard, M., Barforoush, A.A.: Learning ontologies from natural language texts. *Int. J. Hum.-Comput. Stud.* **60**(1), 17–63 (2004)
22. Velardi, P., Navigli, R., Cucchiarelli, A., Neri, F.: Evaluation of OntoLearn, a methodology for automatic population of domain ontologies. In: *Ontology Learning from Text: Methods, Applications and Evaluation*. IOS Press (2006)
23. John Wilbur, W., Sirotkin, K.: The automatic identification of stop words. *J. Inf. Sci.* **18**(1), 45–55 (1992)

TexT - Text Extractor Tool for Handwritten Document Transcription and Annotation

Anders Hast¹, Per Cullhed², and Ekta Vats¹(✉)

¹ Department of Information Technology, Uppsala University, Uppsala, Sweden
`{anders.hast,ekta.vats}@it.uu.se`

² University Library, Uppsala University, Uppsala, Sweden
`per.cullhed@ub.uu.se`

Abstract. This paper presents a framework for semi-automatic transcription of large-scale historical handwritten documents and proposes a simple user-friendly text extractor tool, *TexT* for transcription. The proposed approach provides a quick and easy transcription of text using computer assisted interactive technique. The algorithm finds multiple occurrences of the marked text on-the-fly using a word spotting system. *TexT* is also capable of performing on-the-fly annotation of handwritten text with automatic generation of ground truth labels, and dynamic adjustment and correction of user generated bounding box annotations with the word being perfectly encapsulated. The user can view the document and the found words in the original form or with background noise removed for easier visualization of transcription results. The effectiveness of *TexT* is demonstrated on an archival manuscript collection from well-known publicly available dataset.

Keywords: Handwritten text recognition · Transcription
Annotation · *TexT* · Word spotting · Historical documents

1 Introduction

When printing was invented in the mid 15th century, a sort of transcription revolution took place all over Europe. Single handwritten texts were transformed into multiple copy books. Although this invention was crucial for the growth of knowledge, the process of writing continued well into the 20th century very much as before, with the help of pen and ink.

A similar media-revolution is taking place right now when modern technology in the form of electronic texts is revolutionizing our reading habits and our media distribution possibilities. One of the most crucial steps for science in this modern media-revolution is the ability to search within texts. Optical Character Recognition (OCR) technology [1–3] has opened up even old printed texts to modern science in an unprecedented way. In libraries, meta-data is no longer the sole entry to collections, electronic content can speak for itself and this also changes library practices. However, the large mass of handwritten texts in our

libraries and archives is still waiting to be transformed into searchable texts. The reason for this is a combination of technical and economic factors. Modern technology does not yet give us the good results of OCR technology, which nowadays can be so successfully applied to printed texts that it is a straightforward part of digitization processes world-wide.

Handwritten text recognition (HTR) [4–8] is an emerging field and can be quite successful in certain circumstances, especially when applied to an even and uniform handwriting, but rarely so for the non-homogeneous handwritten texts that fill our archives. In most cases, manual transcription is still the most common way to produce reliable electronic texts from handwritten texts, but modern technology advances and many projects try to tackle this problem. Manual transcription is typically expensive and prone to human error. The incentives to open up this material to computerized searches is high. The information in archives and library collections world-wide, represent an enormously important source to history and only relatively small parts of it is available as electronic texts.

Semi-automatic transcription of manuscripts typically requires hundreds of already transcribed pages, with thousands of examples of each word, in order to produce a useful transcription of the rest of the text. Due to the time consuming machine learning procedures involved, this is computed as off-line batch jobs overnight [7]. However, this means that if just a dozen pages exist, the transcriber is forced to complete the transcriptions without the help of HTR techniques, unless a similar handwriting style exists. An alternative approach to fast transcription of text with a low cost is using computer assisted interactive techniques.

This paper introduces a simple yet effective text extractor tool, *TexT* for transcription of historical handwritten documents. *TexT* is designed for quick document transcription with the help of user interaction where the system finds multiple occurrences of the marked text on-the-fly using a word spotting system. Other advantages of *TexT* include on-the-fly annotation of handwritten text with automatic generation of ground truth labels, adjustment and correction of user labeled bounding box annotations such that the word perfectly fits inside the rectangle. Nevertheless, the transcribed words are cleaned using filtering methods for background noise removal.

This paper is organized as follows. Sections 2, 3 and 4 discusses various transcription and annotation methods and tools available in literature, and discusses related work on handwritten text transcription. Section 5 explains the proposed text extractor tool *TexT* in detail. Section 6 demonstrate the efficacy of the proposed method with implementation details on well-known historical document dataset. Section 7 concludes the paper.

2 Transcription Methods and Tools

Transcriptions can be made by several different techniques, by reading and typing, typically done by one person interested in using the contents of the documents, as opposed to collective transcription where many individuals make

transcriptions using crowdsourcing techniques. HTR, and dictation, are other techniques that can be used to produce transcriptions. An example of the latter is the war-diary of Sven Blom, a Swedish volunteer in The Foreign Legion during the Great War. The diary is kept in Uppsala University Library and was transcribed by dictation [9].

Due to the labour-intensive task involved in transcriptions, crowdsourcing, a term originally coined by Jeff Howe in Wired Magazine in 2006 [10], has been a useful way of distributing transcription work to many people and therefore it sits at the core of many successful transcription projects. The *Transcribe Bentham* project at the University College of London is often mentioned as an example [11]. Like so many others, *Transcribe Bentham* is built with components from the open-source software *MediaWiki*, also used for the perhaps biggest crowdsourcing project on the planet, *Wikipedia*. *Transcribe Bentham* started in 2010 and has to this date completed approximately 43% of the whole collection [12]. They now collaborate with the READ project [13] and the application Transkribus [14], which can combine HTR with manual transcription.

There are numerous other transcription tools on the Internet. Zooniverse [15], based in Oxford, include transcription as one of their crowdsourcing tasks, among many others. The plugin Scripto [16] is one of the oldest, typically created in an environment close to the history discipline, the Roy Rosenzweig Center for History and New Media at George Mason University. It is also based on *MediaWiki* and can be used as a plugin for *Omeka*, *Wordpress* and *Drupal*. Veele Handen [17] is a Dutch application which offers crowdsourced transcriptions as a tool for archives and libraries wishing to open up their collections. They have recently included progress bars where followers and participants can monitor progress.

This feature is very similar to the Smithsonian Institution and their “Digital Volunteers” [18]. In fact, the Smithsonian Institution can be regarded as one of the pioneers in assigning tasks to volunteers. Already in 1849, soon after the founding of Smithsonian Institution, it’s first secretary, Joseph Henry, was able to initiate a network of some 150 volunteers for weather observations, all over the United States [19]. The “Smithsonian Digital Volunteers” is a very successful transcription application and their Graphics User Interface (GUI) combines a clear topical structure with progress bars and a general layout which has incorporated well-established practices used in proof-reading. The work of volunteer number one, has to be approved by a second volunteer and finally the result needs to be approved by the mother institution, wishing to publish the results on the web. Together with other activities, such as promoting projects via social networks, they have managed to achieve good results, demonstrating the importance of an attractive GUI in crowdsourcing. The topical structure facilitates for the user to find attractive tasks.

Uppsala university library is Sweden’s oldest university library and its manuscript collections consist of approximately four kilometers of handwritten material in letters, diaries, notebooks etc. The handwritten manuscript collections date back 2000 years; from BC till the 21st century. The medieval manuscripts

are plentiful and the 16th to 20th centuries are well represented with many single important collections, such as the correspondence of the Swedish King Gustav III, containing letters from, for example the French Queen Marie Antoinette and the Waller collection of 38000 manuscripts with letters from both Isaac Newton and Charles Darwin. The languages in the collection are also diverse (e.g. Swedish, Arabic, Persian etc.). However, the main languages for this project include Swedish, Latin, German, and French.

Since a few years back it has been possible to publish digitized material in the Alvin platform [20], a repository for cultural heritage materials shared among the universities in Uppsala, Lund and Göteborg, as well as other Swedish libraries and museums. However, as so often is the case, very little of the handwritten material is transcribed. The collection can therefore be accessed only through meta-data and cannot be analyzed by computational means, a problem which may only be tackled by long term and multifaceted strategic planning for producing more handwritten document transcriptions.

As a start, Alvin [20] has been adapted to allow for publishing transcriptions alongside the original manuscripts. One example of this is a transcription made from a testimony of refugees arriving to Sweden in 1945 from the concentration camp in Ravenbrück, kept at Lund University Library [21]. In this case, the transcriptions in textual electronic format (such as PDF) are a result of manual transcription and are open to Google indexing, thus making the original manuscripts searchable on the Internet. However, this is only an example, to open up more texts for use in digital humanities, a combination of HTR technology and manual crowdsourced transcriptions is probably as far as our present technologies admit. This work takes an initiative towards transcription and annotation of huge volumes of historical handwritten documents present in our university library using HTR methods such as word spotting [22].

3 Document Annotation Methods and Tools

Several document image ground truth annotation methods [23,24] and tools [25–32] have been suggested in literature. Problems related to ground truth design, representation and creation are discussed in [33]. However, these methods are not suitable for annotating degraded historical datasets with complex layouts [34]. For example, Pink Panther [25], TrueViz [26], PerfectDoc [27] and PixLabeler [28] work well on simple documents only and perform poorly on historical handwritten document images [35].

A highly configurable document annotation tool GEDI [29] supports multiple functionalities such as merging, splitting and ordering. Aletheia [30] is an advanced tool for accurate and cost effective ground truth generation of large collection of document images. WebGT [31] provides several semi-automatic tools for annotating degraded documents and has gained importance recently. Text Encoder and Annotator (TEA) was proposed in [32] for manuscripts annotation using semantic web technologies. However, these tools require specific system

requirements for configuration and installation. Most of these tools and methods are either not suitable for annotating historical handwritten datasets, or represent ground truths with imprecise and inaccurate bounding boxes [35].

Our previous work [34] takes into account such issues, and proposed a simple method for annotating historical handwritten text on-the-fly. This work employs this annotation method with improvements using word spotting algorithm. A detailed discussion of the annotation tools and methods is out of scope of this paper, and the reader is referred to [34] for a deeper understanding of ground truth annotation methods, and on-the-fly handwritten text annotation in general.

4 Related Work on Handwritten Text Transcription

Manual transcription of historical handwritten documents requires highly skilled experts, and is typically a time consuming process. Manual transcription is clearly not a feasible solution due to large amounts of data waiting to be transcribed. Fully automatic transcription using HTR techniques offers a cost-effective alternative, but often fails in delivering the required level of transcription accuracy [36]. Instead, semi-automatic or semi-supervised transcription methods have gained importance in the recent past [36–40].

The transcription method proposed in [40] uses a computer assisted and interactive HTR technique: CATTI (Computer Assisted Transcription of Text Images) for fast, accurate and low cost transcription. For an input text line image to be transcribed, an iterative interactive process is initiated between the CATTI system and the end-user. The system thus generates successively improved transcription in response to the simple user corrective feedback.

Image and language models from partially supervised data have been adapted in [38] to perform computer assisted handwritten text transcription using HMM-based text image modeling and n-gram language modeling. This method has been recently implemented in GIDOC (Gimp-based Interactive transcription of old text Documents) [41] system prototype where confidence measures are estimated using word graphs that helps users in finding transcription errors.

An active learning based handwritten text transcription method is proposed in [39] that performs a sequential line-by-line transcription of the document, and a continuously re-trained system interacts with the end-user to efficiently transcribe each line.

The performance of CATTI system [40], and the methods proposed in [38] and [39] is dependent upon accurate detection of the text lines in each document page. However, the line detection and extraction in historical handwritten document images is a challenging task, and advanced line detection techniques [42] are required.

In practical scenarios, such methods are not appropriate as a system should ideally accept a full document page as an input and generate full transcription of the words as an output. An end-to-end system for handwritten text transcription is presented in [36,37] that also uses HMM-based text image modeling with interactive computer assisted transcription. The transcription method proposed

in this work addresses these issues and introduces *TexT* for quick transcription of handwritten text using a segmentation-free word spotting algorithm [22]. The following section explains the proposed method and its advantages in detail.

5 *TexT* - Text Extractor Tool

This paper presents a framework for semi-automatic transcription of historical handwritten manuscripts and introduces a simple interactive text extractor tool, *TexT* for transcribing words in textual electronic format. The method is based on the idea of transcribing each unique word only once for the whole document, including annotations such as gender, geographical locations, etc. This will both speed up the tedious work of transcription and also make it less exhausting. Furthermore, an interactive approach is proposed where the system finds other occurrences of the same word on-the-fly using so-called word spotting system [22, 43]. The user simply identifies one occurrence, and while the word is being written by the user, the HTR engine finds other possible occurrences of the same word, which are shown to the user, meanwhile it continues in the background to search other pages. Further, the user helps the HTR engine in marking words that are correctly identified and correcting misclassified words. By marking these words, writing their corresponding letter sequence, and adding annotations, the HTR engine in the meanwhile processes these words and more accurately identifies them, making a better distinction between these two classes of words.

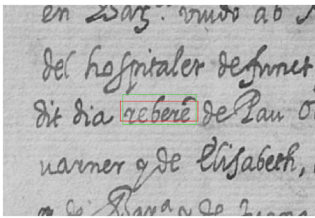
The proposed method inherits features from our previous work [34] and efficiently performs on-the-fly annotation of handwritten text with automatic generation of ground truth labels, and dynamic adjustment and correction of user annotated bounding box labels with perfect encapsulation of the text inside the rectangle. Interestingly, the transcriptions are generated such that the transcribed word contains no added noise from the background or surroundings. This is made possible by the use of two band-pass filtering approach for background noise removal [44]. This is followed by connected components extraction from the word image.

The following features are important parts of the *TexT* project planning:

- A simple yet informative, and user-friendly GUI that may attract users according to well defined topics such as botany, history, theology, diaries, etc.
- A GUI where the user can download the transcription results on-the-fly as they are distributed in the University library digital repository.
- Presence on social networks.
- A ranking system combined with a merit-report for the use of the contributor.
- A proof-reading structure with a first and a second proof-reader and a safe yet quick ingestion mechanism for the repository.
- A graphic illustration of progress for each topic.

- An administration of the application which includes active outreach to find interested audiences, close monitoring of the uploaded content and general advertising of opportunities, news and activities, including events which might give contributors extra value, such as exhibitions and shows of the original material.
- An HTR application, active only in the background, making use of the user input through machine learning and delivering better results based on the user input.

The combination of crowdsourcing and HTR is crucial and, it is believed to be one of the key factors for the *Text* project. Human interaction with AI (artificial intelligence) might be the best way to combine IT-technologies with those interested in contributing to the cultural heritage [45].



reberé

(a) Input document with user marked word (red) and system corrected bounding box (green).

(b) Clean transcribed word *reberé* with background noise removed.

Fig. 1. The user marks a word in the document (in the left), shown in red bounding box. The system finds the best fitting rectangle (in green) to perfectly encapsulate the word. The background noise is removed and the clean transcribed word generated is shown on the right. Figure best viewed in color. (Color figure online)

6 Experimental Framework and Implementation Details

This section emphasize on the overall experimental framework of *Text* along with insight on its implementation details. The proposed framework is tested on the Esposalles dataset [46], a subset of the Barcelona Historical Handwritten Marriages (BH2M) database [47]. BH2M consists of 244 books with information on 550,000 marriages registered between 15th and 19th century. The Esposalles dataset consists of historical handwritten marriages records stored in archives of Barcelona cathedral, written between 1617 and 1619 by a single writer in old Catalan. In total, there are 174 pages handwritten by a single author corresponding to volume 69, out of which 50 pages are selected from 17th century. In future, the ancient manuscripts from the Uppsala University library will be used for further experimentation.

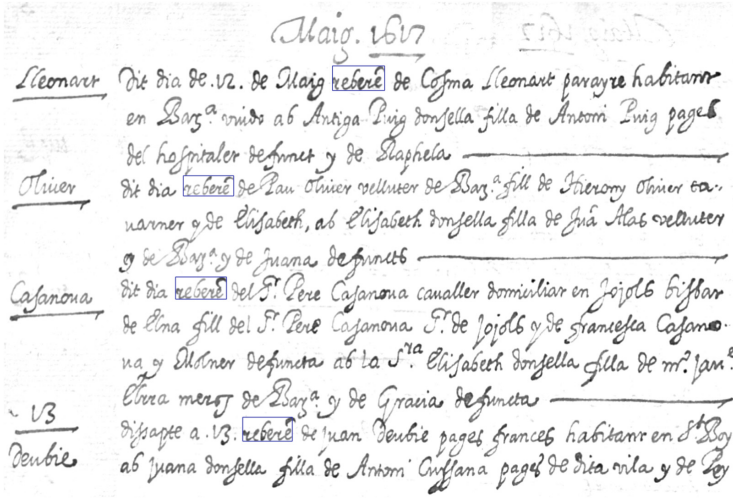


Fig. 2. The result of searching one word marked by the user (for example, reberé), represented using blue bounding box. Figure best viewed in color. (Color figure online)

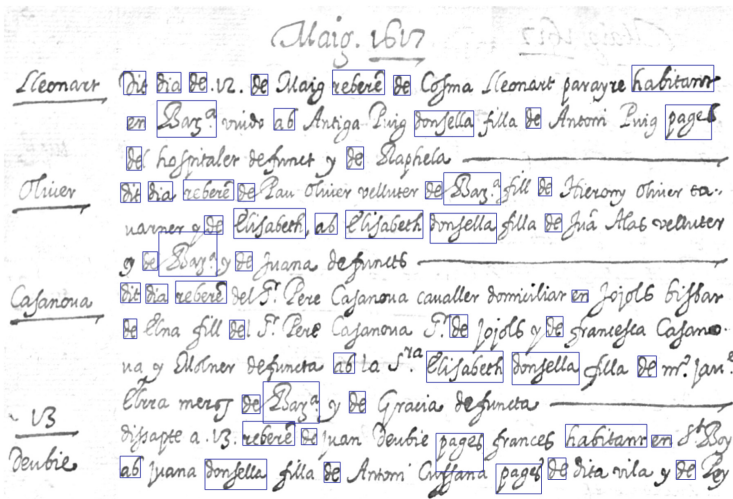


Fig. 3. The transcription can be performed in any order and in this case 11 different words have been marked, and the other occurrences are found automatically. Figure best viewed in color. (Color figure online)

The text transcription method based on word spotting is performed as follows. The system generates a document page query where the user marks a query word with a so called rubber band rectangle. The user marked red bounding box is highlighted in Fig. 1a for a sample word reberé. The system automatically finds the best fitting rectangle to perfectly encapsulate the word, as shown in

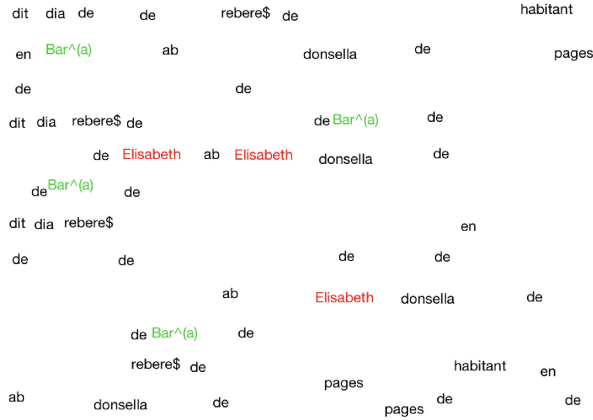


Fig. 4. The ongoing transcription results in words being identified in their corresponding places. In this case, the user has also annotated names and places using different colors. Figure best viewed in color. (Color figure online)

Fig. 1a using green bounding box, and extracts the word. Furthermore, the noise from the background and surroundings is efficiently removed using two band-pass filtering approach in order to make the subsequent search more reliable (see Fig. 1b).

The system starts searching for the word in the document page and the result is shown in Fig. 2. Note that only a cropped part of the document page from the dataset is shown for demonstration. The search is performed while the user inserts the transcribed text together with the annotations. Now the user can let the system learn by clicking on one or several word boxes confirming that they are correctly found. If the system find words that are misclassified, the user can inform the system by clicking a button to switch from correct to incorrect mode, and then selecting the words. While doing this, the system continues to perform word search on other document pages and update the search on the basis of information the system learns from the user (Fig. 4).

The user can select words in any order by marking them once. Figure 3 shows how 11 words have been chosen and the system finds the rest. The corresponding transcription is shown in Fig. 3. In this case, the user has annotated some words as names (highlighted in red) and others as geographical places (highlighted in green). This example of a place represents the abbreviation for the word *Barcelona*.

7 Conclusion and Future Work

The transcription tool *Text* presented in this paper is based on an interactive word spotting system, and lends itself to collaborative work, such as online crowdsourcing for large-scale document transcription. The proposed method can

be further improved using client-server or cloud-based solution to perform transcription without much latency. So far algorithms for word spotting [22] have been developed and a simple experimental framework is proposed to support the transcription approach presented herein.

As future work, we intend to implement a transcription framework on ancient manuscripts from Uppsala University Library that works as follows. Each user can freely mark words, annotate them and also identify words found by the search as correct or incorrect. The major part of the search will be performed on a dedicated computer that splits the work in parallel, making it possible to search even large documents in a few seconds. It can be noted that searching one word in our MATLAB implementation takes about 2s for the example shown in Fig. 2. The word spotting approach used in this work [22] efficiently performs parallel processing such that the search in a single page can be distributed into several processes, and hence making the search much faster. Different learning methods are being evaluated to improve the transcription algorithm. Deep learning techniques can be used only when several hundreds of annotated examples are available for a document, but when starting a transcription of an entirely new document, no such are usually available.

Acknowledgment. This work was supported by the Riksbankens Jubileumsfond (Dnr NHS14-2068:1) and the Swedish strategic research programme eSENCE.

References

1. Mori, S., Nishida, H., Yamada, H.: Optical Character Recognition. Wiley, New York (1999)
2. Govindan, V.K., Shivaprasad, A.P.: Character recognition - a review. *Pattern Recogn.* **23**(7), 671–683 (1990)
3. Blanke, T., Bryant, M., Hedges, M.: Open source optical character recognition for historical research. *J. Doc.* **68**(5), 659–683 (2012)
4. Plamondon, R., Srihari, S.N.: Online and off-line handwriting recognition: a comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(1), 63–84 (2000)
5. Marti, U.V., Bunke, H.: Hidden Markov Models, pp. 65–90. World Scientific Publishing Co., Inc., River Edge (2002)
6. Toselli, A.H., Vidal, E.: Handwritten text recognition results on the Bentham collection with improved classical N-gram-HMM methods. In: Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing, HIP 2015, pp. 15–22. ACM, New York (2015)
7. Espana-Boquera, S., Castro-Bleda, M.J., Gorbe-Moya, J., Zamora-Martinez, F.: Improving offline handwritten text recognition with hybrid HMM/ANN models. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(4), 767–779 (2011)
8. Parvez, M.T., Mahmoud, S.A.: Offline Arabic handwritten text recognition: a survey. *ACM Comput. Sur.* **45**(2), 23:1–23:35 (2013)
9. <http://urn.kb.se/resolve?urn=urn:nbn:se:alvin:portal:record-12537/> (2017)
10. Howe, J.: The rise of crowdsourcing. *Wired Mag.* **14**(6), 1–4 (2006)
11. Moyle, M., Tonra, J., Wallace, V.: Manuscript transcription by crowdsourcing: transcribe bentham. *Liber Q.* **20**(3–4), 347–356 (2011)

12. <http://blogs.ucl.ac.uk/transcribe-bentham/2017/08/21/transcription-update-22-july-to-18-august-2017/> (2017)
13. <http://read.transkribus.eu/>
14. <http://transkribus.eu/Transkribus/>
15. Borne, K., Team, Z.: The zooniverse: a framework for knowledge discovery from citizen science data. In: AGU Fall Meeting Abstracts (2011)
16. <http://scripto.org/>
17. <http://velehanden.nl/> (2017)
18. <http://transcription.si.edu/> (2017)
19. <http://siarchives.si.edu/blog/smithsonian-crowdsourcing-1849/> (2017)
20. <http://www.alvin-portal.org/> (2017)
21. <http://urn.kb.se/resolve?urn=urn:nbn:se:alvin:portal:record-101351/> (2017)
22. Hast, A., Fornés, A.: A segmentation-free handwritten word spotting approach by relaxed feature matching. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS), pp. 150–155. IEEE (2016)
23. Héroux, P., Barbu, E., Adam, S., Trupin, É.: Automatic ground-truth generation for document image analysis and understanding. In: Ninth International Conference on Document Analysis and Recognition, ICDAR 2007, pp. 476–480. IEEE (2007)
24. Pletschacher, S., Antonacopoulos, A.: The page (page analysis and ground-truth elements) format framework. In: 2010 20th International Conference on Pattern Recognition (ICPR), pp. 257–260. IEEE (2010)
25. Yanikoglu, B.A., Vincent, L.: Pink panther: a complete environment for ground-truthing and benchmarking document page segmentation. *Pattern Recogn.* **31**(9), 1191–1204 (1998)
26. Kanungo, T., Lee, C.H., Czorapinski, J., Bella, I.: TRUEVIZ: a groundtruth/metadata editing and visualizing toolkit for OCR. In: Document Recognition and Retrieval VIII, vol. 4307, pp. 1–13. International Society for Optics and Photonics (2000)
27. Yacoub, S., Saxena, V., Sami, S.N.: PerfectDoc: a ground truthing environment for complex documents. In: Proceedings of the Eighth International Conference on Document Analysis and Recognition, pp. 452–456. IEEE (2005)
28. Saund, E., Lin, J., Sarkar, P.: PixLabeler: user interface for pixel-level labeling of elements in document images. In: 10th International Conference on Document Analysis and Recognition, ICDAR 2009, pp. 646–650. IEEE (2009)
29. Doermann, D., Zotkina, E., Li, H.: GEDI - a groundtruthing environment for document images. In: Ninth IAPR International Workshop on Document Analysis Systems (DAS) (2010)
30. Clausner, C., Pletschacher, S., Antonacopoulos, A.: Aletheia - an advanced document layout and text ground-truthing system for production environments. In: 2011 International Conference on Document Analysis and Recognition (ICDAR), pp. 48–52. IEEE (2011)
31. Biller, O., Asi, A., Kedem, K., El-Sana, J., Dinstein, I.: WebGT: an interactive web-based system for historical document ground truth generation. In: 2013 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 305–308. IEEE (2013)
32. Valsecchi, F., Abrate, M., Bacciu, C., Piccini, S., Marchetti, A.: Text encoder and annotator: an all-in-one editor for transcribing and annotating manuscripts with RDF. In: Sack, H., Rizzo, G., Steinmetz, N., Mladenčić, D., Auer, S., Lange, C. (eds.) *ESWC 2016. LNCS*, vol. 9989, pp. 399–407. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47602-5_52

33. Antonacopoulos, A., Karatzas, D., Bridson, D.: Ground truth for layout analysis performance evaluation. In: Bunke, H., Spitz, A.L. (eds.) DAS 2006. LNCS, vol. 3872, pp. 302–311. Springer, Heidelberg (2006). https://doi.org/10.1007/11669487_27
34. Vats, E., Hast, A.: On-the-fly historical handwritten text annotation. In: Proceedings of the 2017 Workshop on Human-Document Interaction (2017, in press)
35. Wei, H., Seuret, M., Liwicki, M., Ingold, R.: The use of Gabor features for semi-automatically generated polygon-based ground truth of historical document images. *Digit. Scholarsh. Humanit.* **32**(1), i134–i149 (2017)
36. Romero, V., Bosch, V., Hernández, C., Vidal, E., Sánchez, J.A.: A historical document handwriting transcription end-to-end system. In: Alexandre, L.A., Salvador Sánchez, J., Rodrigues, J.M.F. (eds.) IbPRIA 2017. LNCS, vol. 10255, pp. 149–157. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58838-4_17
37. Terrades, O.R., Toselli, A.H., Serrano, N., Romero, V., Vidal, E., Juan, A.: Interactive layout analysis and transcription systems for historic handwritten documents. In: 10th ACM Symposium on Document Engineering, pp. 219–222 (2010)
38. Serrano, N., Pérez, D., Sanchis, A., Juan, A.: Adaptation from partially supervised handwritten text transcriptions. In: Proceedings of the 2009 International Conference on Multimodal Interfaces, ICMI-MLMI 2009, pp. 289–292. ACM, New York (2009)
39. Serrano, N., Giménez, A., Sanchis, A., Juan, A.: Active learning strategies for handwritten text transcription. In: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, ICMI-MLMI 2010, pp. 48:1–48:4. ACM, New York (2010)
40. Romero, V., Toselli, A.H., Vidal, E.: Multimodal Interactive Handwritten Text Transcription, vol. 80. World Scientific, Singapore (2012)
41. <https://www.prhlt.upv.es/wp/project/2016/idoc>
42. Bosch, V., Toselli, A.H., Vidal, E.: Semiautomatic text baseline detection in large historical handwritten documents. In: 2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 690–695. IEEE (2014)
43. Giotis, A.P., Sfikas, G., Gatos, B., Nikou, C.: A survey of document image word spotting techniques. *Pattern Recogn.* **68**, 310–332 (2017)
44. Vats, E., Hast, A., Singh, P.: Automatic document image binarization using Bayesian optimization. In: Proceedings of the 2017 Workshop on Historical Document Imaging and Processing. ACM (2017, in press)
45. Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., Horton, J.: The future of crowd work. In: Proceedings of the 2013 conference on Computer Supported Cooperative Work, pp. 1301–1318. ACM (2013)
46. Romero, V., Fornés, A., Serrano, N., Sánchez, J.A., Toselli, A.H., Frinken, V., Vidal, E., Lladós, J.: The ESPOSALLES database: an ancient marriage license corpus for off-line handwriting recognition. *Pattern Recogn.* **46**(6), 1658–1669 (2013)
47. Fernández-Mota, D., Almazán, J., Cirera, N., Fornés, A., Lladós, J.: BH2M: the Barcelona historical, handwritten marriages database. In: 2014 22nd International Conference on Pattern Recognition (ICPR), pp. 256–261. IEEE (2014)

The Distiller Framework: Current State and Future Challenges

Marco Basaldella^(✉), Giuseppe Serra, and Carlo Tasso

Laboratorio di Intelligenza Artificiale, Università degli Studi di Udine,
Via delle Scienze 208, Udine, Italy
{marco.basaldella, giuseppe.serra, carlo.tasso}@uniud.it

Abstract. In 2015, we introduced a novel knowledge extraction framework called the Distiller Framework, with the goal of offering the research community a flexible, multilingual information extraction framework [3]. Two years later, the project has significantly evolved, by supporting more languages and many machine learning algorithms. In this paper we present the current design of the framework and some of its applications.

Keywords: Information extraction · Keyphrase extraction
Named entity recognition

1 Introduction

Today digital document archives contain a tremendous amount of documents of various types, such as books, articles, papers, reports etc. Therefore, there is a urgent demand for adequate tools to semantically process documents in order to support the user needs. Based on this demand, in this paper we present the current state of the Distiller framework, an open source information extraction framework developed in the Artificial Intelligence Laboratory of the University of Udine. The Distiller framework allows annotating any document with linguistic, statistical, semantic or any kind of information.

We present the history of the framework and related research in Sect. 2; in Sect. 3, we describe the design of the framework; then, in Sect. 4, we explain how to download and run the Distiller. In Sect. 5, we briefly present research performed using the Distiller framework in the fields of Keyphrase Extraction and Named Entity Recognition in the biomedical domain. Finally, Sect. 6 presents the challenges that we will have to face in the future for continuing the development of the framework.

2 Related Work

The roots of the Distiller framework are in the Automatic Keyphrase Extraction (herein AKE) system DIKpE [19]. Originally, the system was part of a content recommendation framework, and performed AKE using five features and

heuristically selected weights. Later, [10,11] extended the approach, offering the possibility of inferring keyphrases not contained in the original document and of processing documents in Italian as well. However, the software used in these projects was adapted using a series of ad-hoc solutions, hence becoming difficult to maintain and to further extend with new functionality. For these reasons, we introduced the Distiller framework in [3], with the goal of building a more maintainable system which could be also used for tasks different than AKE.

Other open source KE systems exist in academia. KEA [24], one of the first AKE algorithms, is available online as open source software¹, but the project seems abandoned since 2007. A free implementation of the RAKE [21] algorithm is available online as well², but with little or no possible customization. PKE [8] is an open source³ implementation of many KE algorithms, such as KEA, TopicRank [9], WINGNUS [16] and others. However, it is focused on keyphrase extraction only and it cannot be used for other NLP tasks. The MAUI software⁴ seems the closest system to the Distiller framework, offering an open source implementation of an improved version of the KEA algorithm and algorithms for Named Entity Recognition or Automatic Tagging [15]. However, many of these features are only available buying a commercial license, and the end user is left with no or little possibility of customizing the pipelines.

3 Design

The Distiller framework has been developed in Java 8, due to the robustness of the language, its strong object-oriented paradigm, and due to the availability of a large number of NLP and machine learning tools already available for this language, such as the Stanford CoreNLP library [14], Apache OpenNLP⁵, Weka [22], and others. Moreover, Java gives the possibility of writing wrappers to other software, for added flexibility. Many already available wrappers are developed by the open source community, like e.g. for generic tools like R or Matlab, or for specialized tools like e.g. CRFSuite [17].

The design of the framework is somewhat similar to the Stanford CoreNLP system [14]. In fact, like in Stanford CoreNLP, we offer the possibility to annotate the text with a sequence of `Annotator` objects. When the developer of an information extraction pipeline is working with the Distiller, he will mainly work using the following classes of the framework:

DocumentComponent: this class represents a unit of information within a document. Such unit may be a chapter, a paragraph, a sentence, or just the whole document. It is designed using the Composite pattern [12], where the composite object (sentence, chapter, section...) is represented by the

¹ <http://www.nzdl.org/Kea/download.html>.

² <https://github.com/aneesha/RAKE>.

³ <https://github.com/boudinfl/pke>.

⁴ <https://github.com/zelandiya/maui>.

⁵ <http://opennlp.apache.org>.

`DocumentComposite` class and the smallest component is represented by the `Sentence` class, which is in turn an aggregation of `Tokens`.

Blackboard: this is the class that contains the original document and that will contain all the information produced by the pipeline. It consists of a pointer to the root `DocumentComponent` of the document and a dictionary of `Annotations` that can be filled at according to the specific application considered.

Annotation: the class that represents an annotation. It can be added to the `Blackboard` or any `Annotable` object. Example of `Annotable` objects are any `DocumentComponent`, `Tokens`, `Grams`, etc.

Annotator: an abstract class that has to be extended by any class that produces `Annotations`. An `Annotator` can be a part-of-speech tagger, it can count the occurrences of a word in a document, it can call an external knowledge base (e.g. Wikipedia) to get more information, it can be a machine learning algorithm, and so on.

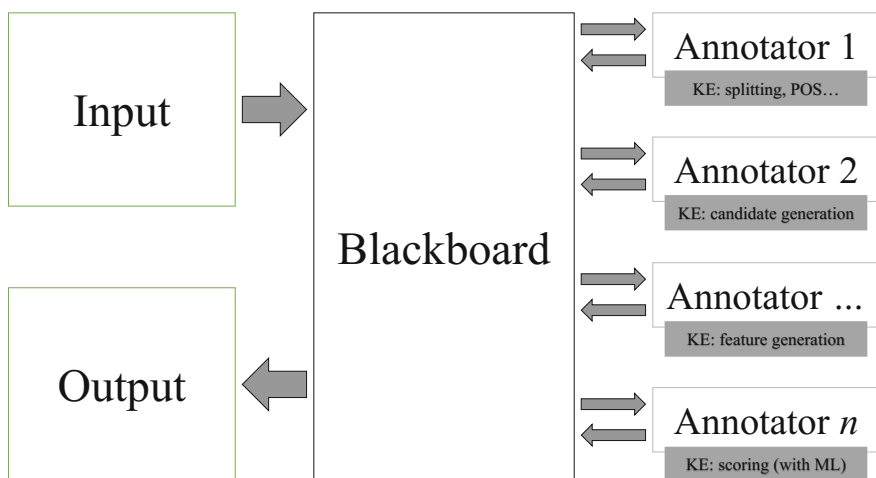


Fig. 1. The high-level architecture of the framework. The workflow is the following: first, the document is written on the blackboard. Then, a sequence of `Annotators` annotate the document, eventually using previously produced annotations. When all annotators finish their job, the produced annotations are returned as output. In this case, we put some example annotators used for Keyphrase Extraction.

Figure 1 shows an example workflow using the Distiller framework, applied to the case of the Keyphrase Extraction task. In this case, the annotators contained in the pipeline will perform the following operations:

1. **Language detection:** first, they detect the language of the document;
2. **Low-level NLP:** then, they perform low-level NLP operations on the document, such as tokenization, part-of-speech tagging, stemming, and so on;

3. **Candidate Generation:** using the information produced in the previous step, they generate the candidate keyphrases that match certain part-of-speech patterns [2, 19];
4. **Candidate Annotation:** they annotate the candidate keyphrases with information from different knowledge domains, such as statistics (e.g. number of occurrences of the candidate, length of the candidate, ...), linguistics (number of nouns in the candidate [19], anaphors that have the candidate as antecedent [2], ...), or from external knowledge (e.g. Wikipedia [3]), and so on;
5. **Candidate Scoring:** they score the candidates using the annotations produced in the previous steps. The score can be calculated using simple, hand-crafted techniques [3, 19] or using machine learning algorithms [2].

Some annotators are already provided out-of-the-box. For example, we provide two wrappers for Stanford CoreNLP and Apache OpenNLP that offer sentence segmentation, word tokenization, and part-of-speech tagging in many languages (see Sect. 5), a wrapper for the Porter’s stemmer algorithm [18], a module that calculates statistical information about n-grams contained in the document, and so on.

4 Obtaining and Running the Distiller Framework

Distiller is available as an open source project under the GPLv2 license, and it is available online at <https://github.com/ailab-uniud/distiller-CORE>.

After building it with Maven, it is possible to run the keyphrase extraction pipeline described in Sect. 3 by writing the following code:

```
String document = ... // load the input document
Distiller d = distiller = DistillerFactory.
    loadFromPackagedXML(“ pipelines/defaultKE.xml”);
Blacboard b = d.distill(document);
Collection<Keyphrase> keyphrase =
    b.getGramsByType(Keyphrase.KEYPHRASE);
```

This code will load the default keyphrase extraction pipeline, run it, and store the results in the `keyphrase` variable. Please note that the `defaultKE` pipeline requires R installed on the system; alternatively, one can run the implementation of [19], called `fastKE`, which does not need any additional software.

5 Applications

Since the beginning of the development of the framework we immediately started to use it in actual research tasks, in order to gain experience about the challenges that developers face in designing tools for the academic world. We actually claim that the flexibility of the Distiller framework, together with its easy customizability, make it an ideal testbed for exploration and for R&D activities.

The following list summarizes the mayor applications and research activities carried out so far in the Artificial Intelligence Laboratory of the University of Udine with the Distiller framework.

Putting More Linguistic and Keyphrase Extraction

We successfully used the Distiller framework to demonstrate the possibility of extracting better keyphrases using more linguistic knowledge than in the classic statistics based approaches [2]. In particular, we exploited the field of Anaphora Resolution (herein AR), obtaining an improvement in performance when adding AR-based information to the KE task. To obtain this result, we used the AR capabilities of the Stanford CoreNLP library to develop two pipelines: one pipeline that replaced anaphors with their antecedents, and one that used AR-based `Annotators` and `Annotations` to score the keyphrases.

Multilinguality and Keyphrase Extraction

We implemented a five-language keyphrase extraction pipeline, capable to process documents in English, Arabic, Portuguese, Romanian and Italian [5]. However, we had training data only for English and Arabic. Thus, to prove the effectiveness of our approach, we trained a machine learning algorithm over the two languages for which we had training data and tested it on custom collected datasets in all five languages. The results show that, even if trained and tested over different languages, the statistical approach of keyphrase extraction is still effective.

Entity Recognition in the Biomedical Domain

We used the Distiller for entity recognition and linking in the biomedical domain as well [4], demonstrating its flexibility. In this work, Distiller was used along OntoGene [20], a text mining framework developed by the University of Zurich, in order to build an hybrid dictionary based - machine learning system for detection and linking of technical terms in the biomedical domain. The system, based on Conditional Random Fields, obtained promising results on the CRAFT corpus, with increased F1-Score when compared to the current state-of-the art systems.

6 Conclusions and Future Work

In the last years, Deep Learning (abbrv. DL) techniques are attacking “classical” machine learning approaches in many fields, outperforming them in many tasks. For example, in the Machine Translation domain, the WMT 2016 task saw a surge of Neural Machine Translation systems, which vastly outperformed the syntax-based systems presented in the previous edition [6, 7]. The same happened in the ImageNet competition, where the introduction of DL techniques

brought the error rate down to 3,6% from the previous, pre-“Deep Learning era” 26,1% state-of-the art [13]. DL approaches also improved the state of the art in many other fields, such as speech recognition and image segmentation, and are currently regarded of obtaining “superhuman” performance in traffic sign classification [13].

DL techniques have been developed also for AKE [25], Named Entity Recognition [23], and many other NLP tasks with very promising results. This will prove a challenge for systems designed to be knowledge-based like the Distiller framework. However, there are tasks where the “classic” knowledge-based approach is still dominant, as NER and Concept Recognition in specialized fields, due to the need of binding concepts to specialized ontologies [4]. In these cases, where algorithms need to take into account rare words that DL models often fail to recognize, we believe systems like the Distiller still has much to offer to the research community. In addition we believe that, due to the extensible framework architecture of Distiller, our system will continue to be useful in the future, e.g. by integrating deep learning libraries inside it. For example, the popular Tensorflow [1] library offers Java APIs, so it would be easy to integrate it in our system. In the future, our goal is to use the Distiller framework to develop an hybrid approach for AKE, which takes into account both the “classic” supervised features, along with the new, DL based techniques.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015). <http://tensorflow.org/>
2. Basaldella, M., Chiaradia, G., Tasso, C.: Evaluating anaphora and coreference resolution to improve automatic keyphrase extraction. In: COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, Osaka, Japan, 11–16 December 2016, pp. 804–814 (2016)
3. Basaldella, M., De Nart, D., Tasso, C.: Introducing distiller: a unifying framework for knowledge extraction. In: Proceedings of 1st AI*IA Workshop on Intelligent Techniques At Libraries and Archives co-located with XIV Conference of the Italian Association for Artificial Intelligence (AI*IA 2015). Associazione Italiana per l’Intelligenza Artificiale (2015)
4. Basaldella, M., Furrer, L., Colic, N., Ellendorff, T., Tasso, C., Rinaldi, F.: Using a hybrid approach for entity recognition in the biomedical domain. In: Proceedings of the 7th International Symposium on Semantic Mining in Biomedicine, SMBM 2016, Potsdam, Germany, 4–5 August 2016, pp. 11–19 (2016)
5. Basaldella, M., Helmy, M., Antolli, E., Popescu, M.H., Serra, G., Tasso, C.: Exploiting and evaluating a supervised, multilanguage keyphrase extraction pipeline for under-resourced languages. In: Recent Advances in Natural Language Processing 2017 (RANLP 2017), Varna (Bulgaria), 4–6 September 2017

6. Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., Zampieri, M.: Findings of the 2016 conference on machine translation. In: Proceedings of the First Conference on Machine Translation, pp. 131–198. Association for Computational Linguistics, Berlin, August 2016
7. Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., Turchi, M.: Findings of the 2015 workshop on statistical machine translation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation, pp. 1–46. Association for Computational Linguistics, Lisbon, September 2015
8. Boudin, F.: pke: an open source python-based keyphrase extraction toolkit. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations. pp. 69–73. The COLING 2016 Organizing Committee, Osaka, December 2016
9. Bougouin, A., Boudin, F., Daille, B.: Topicrank: graph-based topic ranking for keyphrase extraction. In: Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, 14–18 October 2013, pp. 543–551 (2013)
10. De Nart, D., Tasso, C.: A domain independent double layered approach to keyphrase generation. In: WEBIST 2014 - Proceedings of the 10th International Conference on Web Information Systems and Technologies, pp. 305–312. SciTePress (2014)
11. Degl’Innocenti, D., De Nart, D., Tasso, C.: A new multi-lingual knowledge-base approach to keyphrase extraction for the Italian language. In: Proceedings of the 6th International Conference on Knowledge Discovery and Information Retrieval, pp. 78–85. SciTePress (2014)
12. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns: Elements of Reusable Object-oriented Software. Addison-Wesley Longman Publishing Co., Inc., Boston (1995)
13. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016). <http://www.deeplearningbook.org>
14. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations, pp. 55–60 (2014)
15. Medelyan, O., Frank, E., Witten, I.H.: Human-competitive tagging using automatic keyphrase extraction. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, vol. 3, pp. 1318–1327. Association for Computational Linguistics, Stroudsburg (2009)
16. Nguyen, T.D., Luong, M.: WINGNUS: keyphrase extraction utilizing document logical structure. In: Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, 15–16 July 2010, pp. 166–169 (2010). <http://aclweb.org/anthology/S/S10/S10-1035.pdf>
17. Okazaki, N.: Crfsuite: a fast implementation of conditional random fields (crfs) (2007). <http://www.chokkan.org/software/crfsuite/>
18. Porter, M.F.: An Algorithm for suffix stripping. In: Readings in Information Retrieval, pp. 313–316. Morgan Kaufmann Publishers Inc., San Francisco (1997)
19. Pudota, N., Dattolo, A., Baruzzo, A., Ferrara, F., Tasso, C.: Automatic keyphrase extraction and ontology mining for content-based tag recommendation. *Int. J. Intell. Syst.* **25**(12), 1158–1186 (2010)

20. Rinaldi, F.: The ontogene system: an advanced information extraction application for biological literature. *EMBnet.journal* **18**(B) (2012)
21. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. In: *Text Mining*, pp. 1–20 (2010)
22. Russell, I., Markov, Z.: An introduction to the weka data mining system (abstract only). In: *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*, Seattle, WA, USA, 8–11 March 2017, p. 742 (2017)
23. dos Santos, C., Guimarães, V.: Boosting named entity recognition with neural character embeddings. In: *Proceedings of the Fifth Named Entity Workshop*, pp. 25–33. Association for Computational Linguistics, Beijing, July 2015
24. Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G.: Kea: practical automatic keyphrase extraction. In: *Proceedings of the Fourth ACM Conference on Digital Libraries*, pp. 254–255. ACM (1999)
25. Zhang, Q., Wang, Y., Gong, Y., Huang, X.: Keyphrase extraction using deep recurrent neural networks on Twitter. In: *Proceedings of Conference on Empirical Methods in Natural Language Processing* (2016)

Applications of Duplicate Detection in Music Archives: From Metadata Comparison to Storage Optimisation

The Case of the Belgian Royal Museum for Central Africa

Joren Six^(✉), Federica Bressan, and Marc Leman

IPEM, Ghent University, Miriam Makebaplein 1, Ghent, Belgium
{joren.six,federica.bressan,marc.leman}@ugent.be

Abstract. This work focuses on applications of duplicate detection for managing digital music archives. It aims to make this mature music information retrieval (MIR) technology better known to archivists and provide clear suggestions on how this technology can be used in practice. More specifically applications are discussed to complement meta-data, to link or merge digital music archives, to improve listening experiences and to re-use segmentation data. To illustrate the effectiveness of the technology a case study is explored. The case study identifies duplicates in the archive of the Royal Museum for Central Africa, which mainly contains field recordings of Central Africa. Duplicate detection is done with an existing Open Source acoustic fingerprinter system. In the set, 2.5% of the recordings are duplicates. It is found that meta-data differs dramatically between original and duplicate showing that merging meta-data could improve the quality of descriptions. The case study also shows that duplicates can be identified even if recording speed is not the same for original and duplicate.

Keywords: MIR applications · Documentation · Collaboration
Digital music archives

1 Introduction

Music Information Retrieval (MIR) technologies have a lot of untapped potential in the management of digital music archives. There seems to be several reasons for this. One is that MIR technologies are simply not well known to archivists. Another reason is that it is often unclear how MIR technology can be applied in a digital music archive setting. A third reason is that considerable effort is often needed to transform a potentially promising MIR research prototype into a working solution for archivists as end-users.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-73165-0_10) contains supplementary material, which is available to authorized users.

In this article we focus on duplicate detection. It is an MIR technology that has matured over the last two decades for which there is usable software available. The aim of the is article is to describe several applications for duplicate detection and to encourage the communication about them to the archival community. Some of these applications might not be immediately obvious since duplicate detection is used indirectly to complement meta-data, link or merge archives, improve listening experiences and it has opportunities for segmentation. These applications are grounded in experience with working on the archive of the Royal Museum for Central Africa, a digitised audio archive of which the majority of tracks are field recordings from Central Africa.

2 Duplicate Detection

The problem of duplicate detection is defined as follows:

How to design a system that is able to compare every audio fragment in a set with all other audio in the set to determine if the fragment is either unique or appears multiple times in the complete set. The comparison should be robust against various artefacts.

The artefacts in the definition above include noise of various sources. This includes imperfections introduced during the analog-to-digital (A/D) conversion. Artefacts resulting from mechanical defects, such as clicks from gramophone discs or magnetic tape hum. Detecting duplicates should be possible when changes in volume, compression or dynamics are introduced as well.

There is a distinction to be made between *exact*, *near* and *far* duplicates [1]. Exact duplicates contain the exact same information, near duplicates are two tracks with minor differences e.g. a lossless and lossy version of the same audio. Far duplicates are less straightforward. A far duplicate can be an edit where parts are added to the audio – e.g. a radio versus an album edit with a solo added. Live versions or covers of the same song can also be regarded as a far duplicate. A song that samples an original could again be a far duplicate. In this work we focus on duplicates which contain the *same recorded material* from the original. This includes samples and edits but excludes live versions and covers.

The need for duplicate detection is there since, over time, it is almost inevitable that duplicates of the same recording end up in a digitised archive. For example, an original field recording is published on an LP, and both the LP as the original version get digitised and stored in the same lot. It is also not uncommon that an archive contains multiple copies of the same recording because the same live event was captured from two different angles (normally on the side of the parterre and from the orchestra pit), or because before the advent of digital technology, copies of degrading tapes were already being made on other tapes. Last but not least, the chance of duplicates grows exponentially when different archives or audio collections get connected or virtually merged, which is a desirable operation and one of the advantages introduced by the digital technology (see Sect. 2).

From a technical standpoint and using the terminology from [2] a duplicate detector needs to have the following requirements:

- It needs to be capable to mark duplicates without generating false positives or missing true positives. In other words **precision and recall** need to be acceptable.
- It should be capable to operate on large archives. It should be **efficient**. Efficient here means quick when resolving a query and efficient on storage and memory use when building an index.
- Duplicates should be marked as such even if there is noise or the speed is not kept constant. It should be **robust** against various modifications.
- Lookup for short audio fragments should be possible, the algorithm should be **granular**. A resolution of 20 s or less is beneficial.

Once such system is available, several applications are possible (in [1] many of these applications are described as well but, notably, the application of re-use of segmentation boundaries is missing).

Duplicate detection for complementing meta-data. Being aware of duplicates is useful to **check or complement meta-data**. If an item has richer meta-data than a duplicate, the meta-data of the duplicate can be integrated. With a duplicate detection technology conflicting meta-data between an original and a duplicate can be resolved or at least flagged. The problem of conflicting meta-data is especially prevalent in archives with ethnic music where often there are many different spellings of names, places and titles. Naming instruments systematically can also be very challenging.

Duplicate detection to improve the listening experience. When multiple recordings in sequence are marked as exact duplicates, meaning they contain the exact same digital information, this **indicates inefficient storage use**. If they do not contain exactly the same information it is possible that either the same analog carrier was accidentally digitised twice or there are effectively two analogue copies with the same content. To **improve the listening experience** the most qualitative digitised version can be returned if requested, or alternatively to assist philological research all the different versions (variants, witnesses of the archetype) can be returned.

Duplicate detection for segmentation. It potentially solves **segmentation** issues. When an LP is digitised as one long recording and the same material has already been segmented in another digitisation effort, the segmentation boundaries can be reused. Also duplicate detection allows to identify when different segmentation boundaries are used. Perhaps an item was not segmented in one digitisation effort while a partial duplicate is split and has an extra meta-data item – e.g. an extra title. Duplicated detection allows re-use of segmentation boundaries or, at the bare minimum, indicate segmentation discrepancies.

Duplicate detection for merging archives. Technology makes it possible to **merge or link digital archives** from different sources – e.g. the creation of a single point of access to documentation from different institutions concerning a special subject; the implementation of the “virtual re-unification” of collections and holdings from a single original location or creator now widely scattered [3, p. 11]. More and more digital music archives ‘islands’ are bridged by efforts such as Europeana Sounds. Europeana Sounds is a European effort to standardise meta-data and link digital music archives. The EuropeanaConnect/DISMARC Audio Aggregation Platform provides this link and could definitely benefit from duplicate detection technology and provide a view on unique material.

If duplicates are found in one of these merged archives, all previous duplicate detection applications come into play as well. How similar is the meta-data between original and duplicate? How large is the difference in audio quality? Are both original and duplicate segmented similarly or is there a discrepancy?

2.1 Robustness to Speed Change

Duplicate detection robust to speed changes has an important added value. When playback (or recording) speed changes from analogue carriers, both tempo and pitch change accordingly. Most people are familiar with the effect of playing a 33 rpm LP at 45 rpm. But the problem with historic archives and analogue carriers is more subtle: the speed at which the tape gets digitised might not match the original recording speed, impacting the resulting pitch. Often it is impossible to predict with reasonable precision when the recording device was defective, inadequately operated, or when the portable recorder was slowly running out of battery.

So not only it is nearly impossible to make a good estimation of the original non-standard recording speed, but it might not be a constant speed at all, it could actually fluctuate ‘around’ a standard speed. This is also a problem with wax cylinders, where there are numerous speed indications but they are not systematically used – if indications are present at all. In the impossibility to solve this problem with exact precision, a viable approach, balancing out technical needs and philological requirements, is normally to transfer the audio information at standard speed with state-of-the-art perfectly calibrated machinery. The precision of the A/D transfer system in a way compensates for the uncertainty of the source materials. We still obtain potentially sped-up or slowed-down versions of the recording, but when the original context in which the recording was produced can be reconstructed, it is possible to add and subtract quantities from the digitised version because that is exactly known (and its parameters ought to be documented in the preservation meta-data). If the playback speed during transfer is tampered, adapted, guessed, anything that results in a non-standard behaviour in the attempt of matching the original recording speed, will do nothing but add uncertainty to uncertainty, imprecision to imprecision.

An additional reason to digitise historical audio recordings at standard speed and with state-of-the-art perfectly calibrated machinery, is that by doing so, the

archive master [4] will preserve the information on the fluctuations of the original. If we are to “save history, not rewrite it” [5], then our desire to “improve” the quality of the recording during the process of A/D conversion should be held back. Noises and imperfections present in the source carrier bear witness to its history of transmission, and as such constitute part of the historical document. Removing or altering any of these elements violates basic philological principles [6] that should be assumed in any act of digitisation which has the ambition to be culturally significant. The output of a process where sources have been altered (with good or bad intention, consciously or unconsciously, intentionally or unintentionally, or without documenting the interventions) is a *corpus* that is not authentic, unreliable and for all intents and purposes useless for scientific studies. Therefore, in the light of what has been said so far, the problem of speed fluctuation is structural and endemic in historical analogue sound archives, and cannot be easily dismissed. Hence the crucial importance of algorithms that treat this type of material to consider this problem and operate accordingly.

3 Acoustic Fingerprinting

Some possible applications of duplicate detection have been presented in the previous section, now we see how they can be put into practice. It is clear that naively comparing every audio fragment – e.g. every five seconds – with all other audio in an archive quickly becomes impractical, especially for medium-to-large size archives. Adding robustness to speed changes to this naive approach makes it downright impossible. An efficient alternative is needed and this is where *acoustic fingerprinting techniques* comes into play, a well researched MIR topic.

The aim of acoustic fingerprinting is to generate a small representation of an audio signal that can be used to reliably identify identical, or recognise similar, audio signals in a large set of reference audio. One of the main challenges is to design a system so that the reference database can grow to contain millions of entries. Over the years several efficient acoustic fingerprinting methods have been introduced [1, 7–9]. These methods perform well, even with degraded audio quality and with industrial sized reference databases. However, these systems are not designed to handle duplicate detection when speed is changed between the original and duplicate. For this end, fingerprinting system robust against speed changes are desired.

Some fingerprinting systems have been developed that take pitch-shifts into account [10–12] without allowing time-scale modification. Others are designed to handle both pitch and time-scale modification [13, 14]. The system by [13] employs an image processing algorithm on an auditory image to counter time-scale modification and pitch-shifts. Unfortunately, the system is computationally expensive, it iterates the whole database to find a match. The system by [14] allows extreme pitch-shifting and time-stretching, but has the same problem.

The ideas behind both [15, 16] allow efficient duplicate detection robust to speed changes. The systems are built mainly with recognition of original tracks in DJ-sets in mind. Tracks used in DJ-sets are manipulated in various ways

and often speed is changed as well. The problem translates almost directly to duplicate detection for archives. The respective research articles show that these systems are efficient and able to recognise audio with a $\pm 30\%$ speed change.

Only [15] seems directly applicable in practice since it is the only system for which there is runnable software and documentation available. It can be downloaded from <http://panako.be> and has been tested with datasets containing tens of thousands of tracks on a single computer. The output is data about duplicates: which items are present more than once, together with time offsets.

The idea behind Panako is relatively simple. Audio enters the system and is transformed into a spectral representation. In the spectral domain peaks are identified. Some heuristics are used to detect only salient, identifiable peaks and ignore spectral peaks in areas with equal energy – e.g. silent parts. Once peaks are identified, these are bundled to form triplets. Valid triplets only use peaks that are near both in frequency as in time. For performance reasons a peak is also only used in a limited number of triplets. These triplets are the fingerprints that are hashed and stored and ultimately queried for matches.

Exact hashing makes lookup fast but needs to be done diligently to allow retrieval of audio with modified speed. A fingerprint together with a fingerprint extracted from the same audio but with modified speed can be seen in Fig. 1. While absolute values regarding time change, ratios remain the same: $\frac{\Delta t_1}{\Delta t_2} = \frac{\Delta t'_1}{\Delta t'_2}$. The same holds true for the frequency ratios. This information is used in a hash. Next to the hash, the identifier of the audio is stored together with the start time of the first spectral peak.

Lookup follows a similar procedure: fingerprints are extracted and hashes are formed. Matching hashes from the database are returned and these lists are processed. If the list contains an audio identifier multiple times and the start times of the matching fingerprints align in time accounting for an optional linear

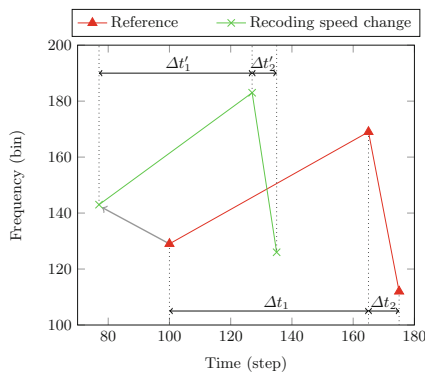


Fig. 1. The effect of speed modification on a fingerprint. It shows a single fingerprint extracted from reference audio ($\text{---}\blacktriangle\text{---}$) and the same fingerprint extracted from audio after recording speed modification ($\text{---}\times\text{---}$).

scaling factor then a match is found. The linear time scaling factor is returned together with the match. An implementation of this system was used in the case study.

4 The Sound Archive of the Royal Museum for Central Africa: A Case Study

The Royal Museum for Central Africa, Tervuren, Belgium preserves a large archive with field recordings mainly from Central Africa. The first recordings were made on wax cylinders in the late 19th century and later on all kinds of analogue carriers were used from various types of gramophone discs to sonofil. During a digitisation project called DEKKMMA (digitisation of the Ethnomusiological Sound Archive of the Royal Museum for Central Africa) [17] the recordings were digitised. Due to its history and size it is reasonable to expect that duplicates are present in the collection. In this case study we want to identify the duplicates, quantify the similarity in meta-data between duplicates and report the number of duplicates with modified speed. Here it is not the aim improve the data itself, this requires specialists with deep knowledge on the archive to resolve or explain (meta-data) conflicts: we mainly want to illustrate the practical use of duplicate detection.

With the Panako [15] fingerprints of 35,306 recordings of the archive were extracted. With the default parameters of Panako this resulted in an index of 65 million fingerprints for 10 million seconds of audio or 6.5 fingerprints per second. After indexing, each recording was split into pieces of 25 s with 5 s overlap, this means a granularity of 20 s. Each of those pieces ($10,000,000 \text{ s} / 20 \text{ s} = 500,000$ items) was compared with the index and resulted in a match with itself and potentially one or more duplicates. After filtering out identical matches, 4,940 fragments of 25 s were found to be duplicates. The duplicate fragments originated from 887 unique recordings. This means that 887 recordings (2.5%) were found to be (partial) duplicates. Thanks to the efficient algorithm, this whole process requires only modest computational power. It was performed on an Intel Core2 Quad CPU Q9650 @ 3.00 GHz, with 8 GB RAM, introduced in 2009.

Due to the nature of the collection, some duplicates were expected. In some cases the collection contains both the digitised version of a complete side of an analogue carrier as well as segmented recordings. Eighty duplicates could be potentially be explained in this way thanks to similarities in the recording identifier. In the collection recordings have an identifier that follows a scheme `collection name.year.collection identifier.subidentifier-track`. If a track identifier contains A or B it refers to a side of an analog carrier (cassette or gramophone disc). The duplicate pair `MR.1979.7.1-A1` and `MR.1979.7.1-A6` suggest that A1 contains the complete side and A6 is track 6 on that side. The following duplicate pair suggests that the same side of a carrier has been digitised twice but stored with two identifiers: `MR.1974.23.3-A` and `MR.1974.23.3-B`. Unfortunately this means that one side is probably not digitised.

The 800 other duplicates do not have similar identifiers and lack a straightforward explanation. These duplicates must have been accumulated over the years. Potentially duplicates entered in the form of analogue copies in donated collections. It is clear that some do not originate from the same analog carrier when listening to both versions. The supplementary material contains some examples. Next, we compare the meta-data difference between original and duplicate.

4.1 Differences in Meta-data

Since the duplicates originate from the same recorded event, to original and duplicate should have identical or very similar meta-data describing their content. This is unfortunately not the case. In general, meta-data implementation depends on the history of an institution. In this case the older field-recordings are often made by priests or members of the military who did not follow a strict methodology to describe the musical audio and its context. Changes in geographical nomenclature over time, especially in Africa, is also a confounding factor [18]. There is also a large amount of vernacular names for musical instruments. The lamellophone for example is known as Kombi, Kembe, Ekembe, Ikembe Dikembe and Likembe [18] to name only a few variations. On top of that, the majority of the Niger-Congo languages are tonal (Yoruba, Igbo, Ashanti, Ewe) which further limits accurate, consistent description with a western alphabet. These factors, combined with human error in transcribing and digitising information, results in an accumulation of inaccuracies. Figure 2 shows the physical meta-data files. If there are enough duplicates in an archive, duplicate detection can serve as a window on the quality of meta-data in general.

Table 1 show the results of the meta-data analysis. For every duplicate a pair of meta-data elements is retrieved and compared. They are either empty, match exactly or differ. Some pairs match quite well but not exactly. It is clear



(a) Filing cabinet in the museum

KONINKLIJK MUSEUM VOOR MIDDENAFRIKA TEWVUUR MUSEE ROYAL DE LAFRIQUE CENTRALE TEWVUUR Reproductiesrecht R. 01334 / C.O. 018 E-1342-D		Band n° 15, 16, 23 Plaat 14 Keuring			
TECHNISCHE GEGEVENS					
Verzameld door / Datum		Opgenomen door		Fotograf	Foto-archief
Joe Ganssens 21, 7, 75		Joe Ganssens Kifwebe / Hekoma			
Band 23/14 ETNOGRAFISCHE IDENTIFICATIE					
PLAATS			VOLK		FUNKTIE
Land / District	Opv.	Stam.	Group		ontspanning
Bwanda	Professeurs 1	Circoncis 1			
Tut	Bakere	Bibayi			
Kinyarwanda	Tut		Kamananga ka Seligira		

(b) Main part of meta-data on file. Some fields use free, handwritten text (e.g. title) others a pre-defined list which are stamped (e.g. language)

Fig. 2. Meta-data on file

Table 1. Comparison of pairs of meta-data fields for originals and duplicates. The field is either empty, different or exactly the same. Allowing fuzzy matching shows that fields are often similar but not exactly the same.

Field	Empty	Different	Exact match	Fuzzy or exact match
Identifier	0.00%	100.00%	0.00%	0.00%
Year	20.83%	13.29%	65.88%	65.88%
People	21.17%	17.34%	61.49%	64.86%
Country	0.79%	3.15%	96.06%	96.06%
Province	55.52%	5.63%	38.85%	38.85%
Region	52.03%	12.16%	35.81%	37.95%
Place	33.45%	16.67%	49.89%	55.86%
Language	42.34%	8.45%	49.21%	55.74%
Functions	34.12%	25.34%	40.54%	40.54%
Title	42.23%	38.40%	19.37%	30.18%
Collector	10.59%	14.08%	75.34%	86.71%

Table 2. Pairs of titles that match only when using a fuzzy match algorithm.

Original title	Duplicate title
Warrior dance	Warriors dance
Amangbetu Olia	Amangbetu olya
Coming out of walekele	Walekele coming out
Nantoo	Yakubu Nantoo
O ho yi yee yi yee	O ho yi yee yie yee
Enjoy life	Gently enjoy life
Eshidi	Eshidi (man's name)
Green Sahel	The green Sahel
Ngolo kele	Ngolokole

that the title of the original *O ho yi yee yi yee* is very similar to the title of the duplicate *O ho yi yee yie yee*. To capture such similarities as well, a fuzzy string match algorithm based on Sørensen–Dice coefficients is employed. When comparing the title of an original with a duplicate, only 19% match. If fuzzy matches are included 30% match. The table makes clear titles often differ while country is the most stable meta-data field. It also makes clear that the overall quality of the meta-data leaves much to improve. To correctly merge meta-data fields requires specialist knowledge - is it *yie* or *yi* - and individual inspection. This falls outside the scope of this case study (Table 2).

4.2 Speed Modifications

In our dataset only very few items with modified speed have been detected. For 98.8% of the identified duplicates the speed matches exactly between original and duplicate. For the remaining 12 identified duplicates speed is changed in a limited range, from -5% to $+4\%$. These 12 pieces must have multiple analogue carriers in the archive. Perhaps copies were made with recording equipment that was not calibrated; or if the live event was captured from multiple angles, it is possible that the calibration of the original recorders was not consistent. There is a number of reasons why a digitised archive ends up containing copies of the same content at slightly different speeds, but it is normally desirable that the cause for this depends on the attributes of the recordings *before* digitisation, and it is not introduced *during* the digitisation process. Our case study shows that duplicates can be successfully detected even when speed is modified. How this is done is explained in the following section.

5 De-duplication in Practice

In this section, the practical functioning of Panako is described. The Panako acoustic fingerprinting suite is Java software and needs a recent Java Runtime. The Java Runtime and TarsosDSP [19] are the only dependencies for the Panako system, no other software needs to be installed. Java makes the application multi-platform and compatible with most software environments. It has a command-line interface, users are expected to have a basic understanding of their command line environment.

Panako contains a deduplicate command which expects either a list of audio files or a text file that contains the full path of audio files separated by new-lines. This text file approach is more practical on large archives. After running the deduplicate program a text file will contain the full path of duplicate files together with the time at which the duplicate audio was detected.

Several parameters need to be set for a successful de-duplication. The main parameters determine the granularity level, allowed modifications and performance levels. The granularity level determines the size of the audio fragments that are used for de-duplication. If this is set to 20 s instead of 10, then the number of queries is, obviously, halved. If speed is expected to be relatively stable, a parameter can be set to limit the allowed speed change. The performance can be modified by choosing the number of fingerprints that are extracted per second. The parameters determine several trade-offs between query speed, storage size, and retrieval performance. The default parameters should have the system perform reasonably effectively in most cases.

The indirect applications of linking meta-data is dependent on organization of the meta-data of the archive but has some common aspects. First, the audio identifiers of duplicates are arranged in original/duplicate pairs. Subsequently, the meta-data of these pairs is retrieved from the meta-data store (e.g. a relational database system). Finally, the meta-data element pairs are compared and resolved. The last step can use a combination of rules to automatically merge

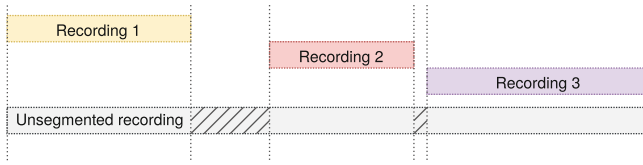


Fig. 3. Reuse of segmentation boundaries. The recording 1, 2 and 3 are found in a long unsegmented track. The segmentation boundaries (dotted lines) can be reused. Some parts in the unsegmented track remain unlabeled (the parts with diagonal lines).

meta-data and manual intervention when a meta-data conflict arises. The manual intervention requires analysis to determine the correct meta-data element for both original and duplicate.

Reuse of segmentation boundaries needs similar custom solutions. However, there are again some commonalities in reuse of boundaries. First, audio identifiers from the segmented set are identified within the unsegmented set resulting in a situation as in Fig. 3. The identified segment boundaries can subsequently be reused. Finally, segments are labeled. Since these tasks are very dependent on file formats, database types, meta-data formats and context in general it is hard to offer a general solutions. This means that while the duplicate detection system is relatively user friendly and ready to use, applying it still needs a software developer but not, and this is crucial, an MIR specialist.

6 Conclusions

In this paper we described possible applications of duplicate detection techniques and presented a practical solution for duplicate detection in an archive of digitised audio of African field recordings. More specifically applications were discussed to complement meta-data, to link or merge digital music archives, to improve listening experiences and to re-use segmentation data. In the case study on the archive of the Royal Museum of Central Africa we were able to show that duplicates can be successfully identified. We have shown that the meta-data in that archive differs significantly between original and duplicate. We have also shown that duplicate detection is robust to speed variations.

The archive used in the case study is probably very similar to many other archives of historic recordings and similar results can be expected. In the case study we have shown that the acoustic fingerprinting software Panako is mature enough for practical application in the field today. We have also given practical instructions on how to use the software. It should also be clear that all music archives can benefit from this technology and we encourage archives to experiment with duplicate detection since only modest computing power is needed even for large collections.

Acknowledgements. This work was partially supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 703937 and partly supported by an FWO Methusalem project titled *Expressive Music Interaction*.

References

1. Orio, N.: Searching and classifying affinities in a web music collection. In: Agosti, M., Bertini, M., Ferilli, S., Marinai, S., Orio, N. (eds.) IRCDL 2016. CCIS, vol. 701, pp. 59–70. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56300-8_6
2. Cano, P., Batlle, E., Kalker, T., Haitsma, J.: A review of audio fingerprinting. *J. VLSI Signal Process.* **41**, 271–284 (2005)
3. IFLA - Audiovisual and Multimedia Section: Guidelines for digitization projects: for collections and holdings in the public domain, particularly those held by libraries and archives. Technical report, International Federation of Library Associations and Institutions (IFLA), Paris, France, March 2002
4. IASA-TC 2004: Guidelines on the Production and Preservation of Digital Objects. IASA Technical Committee (2004)
5. Boston, G.: Safeguarding the Documentary Heritage. A guide to Standards, Recommended Practices and Reference Literature Related to the Preservation of Documents of all kinds. UNESCO (1998)
6. Bressan, F., Canazza, S., Vets, T., Leman, M.: Hermeneutic implications of cultural encoding: a reflection on audio recordings and interactive installation art. In: Agosti, M., Bertini, M., Ferilli, S., Marinai, S., Orio, N. (eds.) IRCDL 2016. CCIS, vol. 701, pp. 47–58. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56300-8_5
7. Wang, A.L.C.: An industrial-strength audio search algorithm. In: Proceedings of the 4th International Symposium on Music Information Retrieval (ISMIR 2003), pp. 7–13 (2003)
8. Haitsma, J., Kalker, T.: A highly robust audio fingerprinting system. In: Proceedings of the 3th International Symposium on Music Information Retrieval (ISMIR 2002) (2002)
9. Ellis, D., Whitman, B., Porter, A.: Echoprint - an open music identification service. In: Proceedings of the 12th International Symposium on Music Information Retrieval (ISMIR 2011) (2011)
10. Fenet, S., Richard, G., Grenier, Y.: A scalable audio fingerprint method with robustness to pitch-shifting. In: Proceedings of the 12th International Symposium on Music Information Retrieval (ISMIR 2011), pp. 121–126 (2011)
11. Bellettini, C., Mazzini, G.: Reliable automatic recognition for pitch-shifted audio. In: Proceedings of 17th International Conference on Computer Communications and Networks (ICCCN 2008), pp. 838–843. IEEE (2008)
12. Ramona, M., Peeters, G.: AudioPrint: an efficient audio fingerprint system based on a novel cost-less synchronization scheme. In: Proceedings of the 2013 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2013), pp. 818–822 (2013)
13. Zhu, B., Li, W., Wang, Z., Xue, X.: A novel audio fingerprinting method robust to time scale modification and pitch shifting. In: Proceedings of the international conference on Multimedia (MM 2010), pp. 987–990. ACM (2010)
14. Malekesmaeili, M., Ward, R.K.: A local fingerprinting approach for audio copy detection. Computing Research Repository (CoRR) abs/1304.0793 (2013)

15. Six, J., Leman, M.: Panako - a scalable acoustic fingerprinting system handling time-scale and pitch modification. In: Proceedings of the 15th ISMIR Conference (ISMIR 2014), pp. 1–6 (2014)
16. Sonnleitner, R., Widmer, G.: Quad-based audio fingerprinting robust to time and frequency scaling. In: Proceedings of the 17th International Conference on Digital Audio Effects (DAFx-2014) (2014)
17. Cornelis, O., De Caluwe, R., Detré, G., Hallez, A., Leman, M., Matthé, T., Moelants, D., Gansemans, J.: Digitisation of the ethnomusicological sound archive of the RMCA. *IASA J.* **26**, 35–44 (2005)
18. Cornelis, O., Lesaffre, M., Moelants, D., Leman, M.: Access to ethnic music: advances and perspectives in content-based music information retrieval. *Sig. Process.* **90**(4), 1008–1031 (2010). Special Section: Ethnic Music Audio Documents: From the Preservation to the Fruition
19. Six, J., Cornelis, O., Leman, M.: TarsosDSP, a real-time audio processing framework in Java. In: Proceedings of the 53rd AES Conference (AES 53rd), The Audio Engineering Society (2014)

Extracting Dependency Relations from Digital Learning Content

Giovanni Adorni¹, Felice Dell’Orletta², Frosina Koceva^{1(✉)}, Ilaria Torre¹,
and Giulia Venturi²

¹ Department of Informatics, Bioengineering, Robotics and Systems Engineering,
University of Genoa, Genoa, Italy

{giovanni.adorni,ilaria.torre}@unige.it, frosina.koceva@edu.unige.it

² Istituto di Linguistica Computazionale Antonio Zampolli (ILCCNR), Pisa, Italy
{felice.dellorletta,giulia.venturi}@ilc.cnr.it

Abstract. Digital Libraries present tremendous potential for developing e-learning applications, such as text comprehension and question-answering tools. A way to build this kind of tools is structuring the digital content into relevant concepts and dependency relations among them. While the literature offers several approaches for the former, the identification of dependencies, and specifically of prerequisite relations, is still an open issue. We present an approach to manage this task.

Keywords: Prerequisite relationship · Concept extraction
Graph mining

1 Introduction

The 21th century is marked by the exponential growth of data and of digital contents. Digital libraries evolved from static storage and retrieval platforms to dynamic services to explore, exchange and share information and knowledge.

In this paper, our focus is on the potential role of digital libraries for education. The idea is that digital resources can not only be explored and shared but they can be coupled with services that support learning processes. This usually requires that content is extracted, structured and enriched with annotations. Since the objective is supporting learning, the extraction of relevant concepts has to be complemented with the identification of prerequisite relations among these concepts. This enables the building of services that, for example, enable to find pieces of knowledge in the text and to extract also the related propaedeutic concepts and resources that allow such information to be properly understood (prerequisite relations).

Manual annotation is of course the most effective approach, but it is time consuming and requires experts knowledge. Therefore, a challenge is the automatic learning of the knowledge structure of the content.

While several methods exist (e.g., [1, 3]) to face the issue of concept extraction, the identification of prerequisite relations among concepts is still an open research problem. In this paper we present methods and approaches for facing this issue.

2 Research Issue and Background

The two main tasks for automatic concept map building are the concept extraction and the relations identification between concepts [7]. Even though there is a long-standing interest since at least 1971 Gagné’s work on learning hierarchies [6], identifying prerequisite relations among concepts is an open issue.

The prerequisite relation between two concepts A and B is a dependency relation which represents what a learner must know/study (concept A), before approaching concept B. Thus, A is a propaedeutic concept, i.e. a requirement, for B and the learner should first understand A in order to understand B.

The prerequisite relation can represent a hyponymy or meronymy relation in the case where the hyponym/meronym concept is going to be further in-depth studied and therefor is itself a prerequisite to another concepts. The prerequisite relation usually requires experts to be evaluated since its semantics can be properly evaluated only by considering the whole graph and the learning goal.

Notation. In the following we provide the conventions and definitions that will be used along the paper. We define a document D as a textual resource. The output of the concept extraction is the terminology $T \in D$ with $t \in T$, where t is a domain-specific term, composed of one or more words (single nominal terms or complex nominal structures with modifiers). For each term, the process returns also its relevance $r = [0, 1]$ (see Sect. 3 for definition).

When D is structured into parts, sections (S), the output of the concept extraction can be $T \in D$ and $T \in S$ according to the needs. Subsections are managed as Sections. Thus we have concept-document and concept-section relationships. We denote these relationships as relevance functions $F(\cdot, \cdot)$ which take the concept and D/S as arguments and have the relevance r as output.

The final output of concepts and prerequisite relations extraction is a concept graph G . Similarly to [10], we represent G as a set of triples in the form $G = \{(t_1, t_2, p) | t_1, t_2 \in T, 0 \leq p \leq 1\}$, where p is the prerequisite relationship and can take a value from 0 to 1, indicating the strength of the prerequisite relation between t_1 and t_2 (where t_1 is prerequisite of t_2).

Term appearance in section is defined as a pair (t_i, s_j) , $t_i \in T$ and $s_j \in S$.

3 Concept Extraction

Our approach to the identification of prerequisite relations was tested on the handbook entitled *Computer Science: An Overview: Global Edition*, G. Brookshear and D. Brylow, Pearson 2015. In order to identify relevant concepts within the considered book, we exploited Text-To-Knowledge (T2K²) [3], a software platform developed at the Institute of Computational Linguistics “A. Zampolli” of the CNR in Pisa. T2K² relies on a battery of tools for Natural Language Processing, statistical text analysis and machine learning which are dynamically integrated to provide an accurate representation of the linguistic information and of the domain-specific content of multilingual text corpora. T2K² encompasses two main sets of modules, respectively devoted to carry

out the linguistic pre-processing of the acquisition corpus and to extract and organize the domain knowledge contained in the linguistically annotated texts. Each section of the considered handbook was automatically enriched (i.e. annotated) with linguistic information at increasingly complex levels of analysis, represented by sentence splitting, tokenization, Part-Of-Speech tagging and lemmatization. According to the methodology described in [2], the automatically POS-tagged and lemmatized input text is searched for candidate domain-specific terms denoting domain entities expressed by either single nominal terms (e.g. *internet*, *network*, *software*) or complex nominal structures with modifiers (typically, adjectival and prepositional modifiers), where the latter are retrieved on the basis of a set of POS patterns (e.g. adjective + noun, noun + preposition + noun) encoding morpho-syntactic templates for multi-word terms (e.g. *Internet Protocol*, *eXtensible Markup Language*, *client/server model*). The domain relevance of both single and multi-word terms t included in the extracted list T is weighted on the basis of the C-NC Value [5] aimed at assessing how much a term is likely to be conceptually independent from the context in which it appears. Accordingly, a higher C-NC rank is assigned to those multi-word terms that are more relevant for the domain of the document collection in input. The extracted domain-specific entities are organized according to co-occurrence relations, i.e., relations between entities co-occurring within the same context. The relevance of relations is weighted using the log-likelihood metric for binomial distributions as defined by [4]. According to this metric, for example, the term *Internet* is strongly related with *Internet Protocol addresses*, *Simple Mail Transfer protocol*, *message*, etc. The extracted relations between terms can be visualized in a ‘knowledge graph’ which can be exploited in a number of graph analyses. M1 in the next section is based on the knowledge graph.

4 Prerequisite Relationship Identification

In this paper we propose two methods for identifying candidate prerequisite relationships (t_1, t_2, p) , with $p \in [0, 1]$. The underlying principles are:

- Co-occurrence of two concepts is a necessary but not sufficient condition to identify the prerequisite relation. The principle can be extended from the sentence level to a section level.
- Temporal occurrence of terms and/or sections are taken into account to identify the direction of prerequisite relation, with different granularities.

Since the methods exploit these principles in different ways, they are designed to be finally combined in order to exploit the benefits of both the approaches.

Method 1 (M1) is based on temporal order and co-occurrence of terms. Steps:

- Building a list L of terms $t \in T$ ordered according to their temporal appearance in D where the term t has the first significant density (which can be compute with different methods, e.g. Burst Analysis).
- Transforming the undirected knowledge graph from Sect. 3 generated with log-likelihood metric into a directed graph G_1 , where direction is derived from the ordered list of terms L .

Result: Candidate triples for prerequisite relations are the adjacent terms in G_1 (Fig. 1). The G_1 graph is represented as a $n \times n$ matrix M_1 , with $n = |T|$. Each element t_{ij} represents the weight p of the prerequisite relationship between terms t_i and t_j , with $p = [0, 1]$.

The strength of relationship p can be defined using different approaches as: NLP analysis, Lexical pattern and other heuristics.

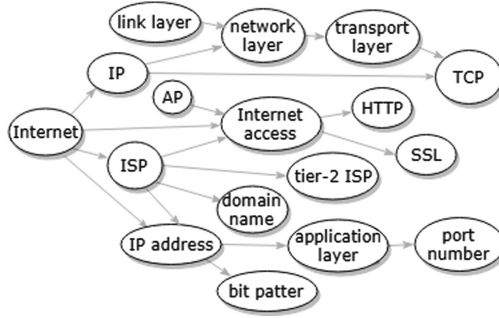


Fig. 1. Method 1 - Examples of candidate prerequisite relations

Method 2 (M2) is based on text structure D/S (Table of Content): The goal of this approach is to identify, for each term, the cluster of terms that are likely or unlikely to be in prerequisite relationship with the term. $TOC(s_i, s_j)$ represents the order \prec of section i and section j , where $s_i, s_j \in S$. The application of the method is represented in the examples in Fig. 2. Steps:

- For each term $t \in T$, identifying the section s_i where the relevance function $F(t, q)$ has max value (i.e., identifying the section where the term has the higher relevance in the document); the assumption is that a concept is explained where it has maximum relevance.
- For each (t_v, s_i) , where $v \neq u$, identifying the section s_j where the relevance function $F(t_v, s_j)$ has max value
 - (i) If $s_j \prec s_i \wedge \nexists (t_u, s_j)$, its unlikely that t_u is a prerequisite of t_v based on the principle that in s_j there should be at least one occurrence of the prerequisite (t_u), see Fig. 2 (i).
 - (ii) If $s_i \prec s_j \wedge \nexists (t_u, s_j)$ is likely that t_v is a prerequisite of t_u , since t_v is explained before t_u and it also co-occurs in s_i , see Fig. 2 (ii).
 - (iii) If $s_j \prec s_i \wedge \exists (t_u, s_j)$ there is some probability that t_v is a prerequisite of t_u , since they could be highly related concepts but not as prerequisite relationship. Similarly, if $s_i \prec s_j \wedge \exists (t_u, s_j)$ there is some probability that t_u is a prerequisite of t_v , for the same reason as in the previous point, see Fig. 2 (iii).
 - (iv) If $s_i = s_j$, thus t_v and t_u co-occur with maximum relevance in the same section, see Fig. 2 (iv), this means that the concepts are highly related but we cannot identify the prerequisite relationship unless further analysis is performed, such as: NLP, Lexical pattern and other heuristics.

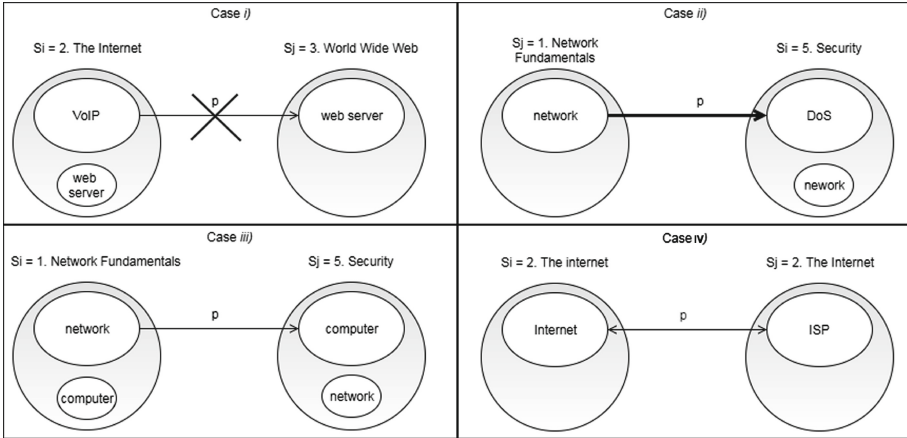


Fig. 2. Method 2 - Examples of candidate prerequisite extraction

Result: The candidate prerequisite relations are represented as a $n \times n$ matrix M_2 , with $n = |T|$. Each element t_{ij} represents the weight p of the prerequisite relationship between terms t_i and t_j , with $p = [0, 1]$. The implementation of the algorithm can apply values of p according to the rules above which can be tuned in order to fit the specific domain.

5 Discussion and Conclusion

In this section we discuss the proposed approach by comparing our methods with related approaches for concept and prerequisite extraction. An approach that exploits textbook internal information (*TOC*) to identify prerequisite relations is adopted in [10], even though they also exploit external knowledge (from Wikipedia) to extract the relevant concepts. Another approach that exploits Wikipedia is described in [8]. The authors define a metric (i.e., refD) that models the relation by measuring how differently two concepts refer to each other. In [9] the authors mine prerequisite relations among MOOC course concepts by defining three main features: semantic (incorporates wikipedia knowledge), contextual (similar to refD [8]) and structural distributional patterns.

Unlike the above cases, our approach exploits only features from the text (co-occurrence, term density, temporal and *TOC* ordering) for concept and prerequisite extraction, without using external knowledge. With respect to [10], while the authors exploit *TOC* title match and order coherence, we identify candidate prerequisite relation by the joint usage of not only *TOC* order (M2) but also the temporal concept density order (M1), thus providing a more granular method. Moreover, while in [10] the information overlap is calculated by using Wikipedia title match and similarity functions, we use concept-section order analysis (M2) to identify three specific cases of concept redundancy of which (ii) identifies prerequisite candidate conceptually similar to refD in [8] where

the sections in our case can be seen as the wikipedia articles in refD. Whereas most of the aforementioned methods for prerequisite extraction result in a concept hierarchy building, i.e. tree structure, the M2 (*iii*) give the bases towards a graph building by adding parallel prerequisite relations.

Enhancement of M1 can be made by introducing metrics based on concept bursting intervals (e.g. [11]) for building the list L. In addition, by analyzing more than one book (with the same subject), both methods can be improved by reducing biases due to the author's subjective choices in structuring the book. We are working on testing the methods and the mentioned enhancements.

Acknowledgements. The authors thank prof. Carlo Tasso for making available Distiller system for concept extraction during the initial experiments of the described methodology.

References

1. Basaldella, M., Chiaradia, G., Tasso, C.: Evaluating anaphora and coreference resolution to improve automatic keyphrase extraction. In: COLING, pp. 804–814 (2016)
2. Bonin, F., Dell'Orletta, F., Venturi, G., Montemagni, S.: A contrastive approach to multi-word term extraction from domain corpora. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (2010)
3. Dell'Orletta, F., Venturi, G., Cimino, A., Montemagni, S.: T2k²: a system for automatically extracting and organizing knowledge from texts. In: Proceedings of 9th International Conference on Language Resources and Evaluation, pp. 2062–2070 (2014)
4. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.* **19**(1), 61–74 (1993)
5. Frantzi, K., Ananiadou, S.: The C-value/NC value domain independent method for multi-word term extraction. *J. NLP* **6**(3), 145–179 (1999)
6. Gagné, R.M.: Learning hierarchies. In: Merrill, M.D. (ed.) *Instructional Design: Readings*, pp. 118–131. Prentice-Hall, Englewood Cliffs (1968, 1971)
7. Kowata, J.H., Cury, D., Boeres, M.: A review of semi-automatic approaches to build concept maps. In: Proceedings of the 4th Conference on Concept Mapping, pp. 40–48 (2010)
8. Liang, C., Wu, Z., Huang, W., Giles, C.L.: Measuring prerequisite relations among concepts. In: EMNLP, pp. 1668–1674 (2015)
9. Pan, L., Li, C., Li, J., Tang, J.: Prerequisite relation learning for concepts in MOOCs. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, Long Papers, vol. 1, pp. 1447–1456 (2017)
10. Wang, S., Ororbia, A., Wu, Z., Williams, K., Liang, C., Pursel, B., Giles, C.L.: Using prerequisites to extract concept maps from textbooks. In: Proceedings of the 25th ACM International Conference on Information and Knowledge Management, pp. 317–326 (2016)
11. Yoon, W.C., Lee, S., Lee, S.: Burst analysis of text document for automatic concept map creation. In: Ali, M., Pan, J.-S., Chen, S.-M., Horng, M.-F. (eds.) *IEA/AIE 2014. LNCS (LNAI)*, vol. 8482, pp. 407–416. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07467-2_43

Annote: A Serious Game for Medical Students to Approach Lesion Skin Images of a Digital Library

Fabrizio Balducci^(✉)

Dipartimento di Ingegneria “Enzo Ferrari”,
Università degli Studi di Modena e Reggio Emilia,
Via Vivarelli 10, 41125 Modena, MO, Italy
fabrizio.balducci@unimore.it

Abstract. Nowadays it is claimed that one method to learn how to execute a task is to present it as a gaming activity: in this way a teacher can offer a safe and controlled environment for learners also arousing excitement and engagement. In this work we present the design of the serious game ‘Annote’, to exploit a medical digital library with the aim to help dermatologists to teach students how to approach the examination of skin lesion images to prevent melanomas.

Keywords: Education · Learning · Serious game · Skin images
Gamification

1 Introduction

The *Gamification* process consists in the application of game-design elements and principles in non-game contexts [16]: it uses the game mechanics to improve skills and knowledge of a subject, also enhancing its engagement and excitement while performing a task that usually does not provides them. Referring to the Csíkszentmihályi [12, 19] and Chen [8] studies the sense of fun is strictly connected with the *Flow theory* characterized by the constant steady and balance between the *challenge* offered to gamers and the *skills* developed while facing them: in [5, 11, 13] are studies about video semantic recognition while the evaluation of affective states and moods are in [3].

An *Exergame* identifies games that are also a form of exercise and involves the creation of a context in which the subject can use certain tools to replicate a series of real movements or tasks: they are used to counteract a sedentary activity, medical rehabilitation and promoting an active lifestyle and they are designed to provide immediate feedback to the player with the possibility of monitoring behaviors and biological parameters.

With the gamification technique it is possible to develop *serious games* which are games designed not only for the pure entertainment: this game genre is focused on the *simulation* feature with pedagogical purpose, by exploiting fun

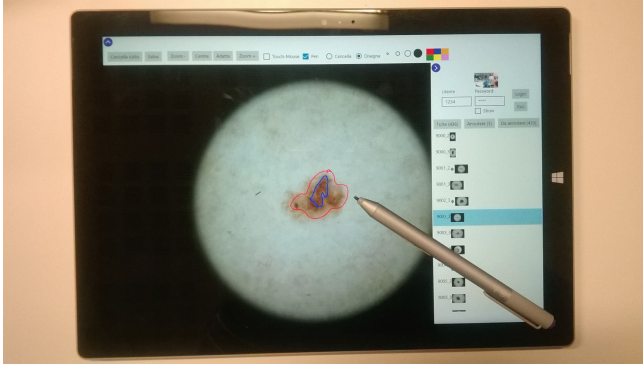


Fig. 1. The original annotation tool developed for Microsoft Surface Pro 3

and competition while used in environments like defense, education [1], scientific exploration, health care [14], emergency management, city politics [17].

The aim of this work is to apply the gamification process to one of the main activity in the daily work of dermatologists with which they make diagnoses and comparisons: the medical image annotation. The paper is organized as follows: Sect. 2 describes the general architecture and the design assumptions taken for the videogame while Sect. 3 shows technical details for the development of the prototype; finally conclusions and future work are illustrated in Sect. 4.

2 The Gamification Process

The idea to develop a serious (video)game for medical teaching (taking also inspiration from exergames) comes from the previous works for the development of an annotation tool [2] (Fig. 1) used by academic dermatologists to annotate skin lesion images and build an integrated medical Digital Library (Fig. 2).

The tool was designed following several interaction principles like *affordance* and *direct manipulation* [20,21]) with the aim to extract a dataset of visual features to use with Machine Learning techniques for a recommendation agent. During a preliminary test session, senior dermatologists and academic interns expressed interest towards innovative learning methods like serious gaming.

The survey of Hamari *et al.* [15] presents ten *motivational affordances* tested on empirical studies about gamification that will guide the design of Annote:

(h1) Score points (h2) Leaderboard (h3) Achievements (h4) Levels (h5) Challenge (h6) Story/Theme (h7) Goals (h8) Feedback (h9) Rewards (h10) Progress

The work of Coltell *et al.* [9] takes up the previous aspects and adds:

(c1) Rules (c2) Safety (c3) Interaction (c4) People (c5) Fantasy (c6) Exploration

The *game objects* is the act of ‘draw strokes’ and the right use of interactive tools: the *repetitiveness* is a learning element that, differently from commercial videogames, reinforces behavior change and progression in performances [7, 22].

The client-server architecture, the tools and the user interface (see Sect. 3) ensures that h8, c1, c2, c3 and c6 are met. A imaginative story (h6 and c5) results hard to introduce if not with the *survival mission* like “save lives as fast as possible” while there is a separate section with a non-interactive tutorial.

The main types of *challenge* offered by the game (c1, h5, h7, h10) are:

1. border challenge (precision): the player has to draw a lesion border annotation that imitates the ‘official’ one (ground-truth) also considering the completion of already begun strokes
2. structures challenge (recognition): the player has to annotate not lesion borders but groups of skin *textures*, *clues* and *patterns* (lines, circles, reticles, ..)
3. time challenge (pressure): a variant of the previous two where the player must annotate respecting a flowing timer for each image
4. lesion classification (quiz): the player looks an image and gives a diagnosis on the severity of a lesion by choosing from a ranked set (Likert scale).

It must be noticed that the ground-truth has been previously made for the digital library by the academics (or by ‘official’ algorithms) so this guarantees the quality and the reliability of an annotation chosen for gaming comparison.

The variety of the game (h4, h5, h7, h10) is also given by the *difficulty* modes:

- the type and amount of images chosen by the teacher for a game session
- the activation of aid/impediment game features.

To expand and diversify the gameplay, forms of *rewards* are introduced (h1, h2, h3, h5, h10, c4):

- power-up or penalty: they grant or steal resources to the player (time, points) or enable/disable features like the available tools (zoom, stroke width/color) or their performances (image resolution, mouse/pen speed)
- points and leaderboard: a centralized classification which emphasizes the desire to improve gaming (and learning) performances between students [18]
- personal profile: customizable, summarizes player informations
- badges and achievements: ‘titles’ that appear in the player profile and in the leaderboard screen; they depend by the number of accomplished tasks and by the amount of gained points.

3 Developing the Serious Game Annote

This videogame allows to draw strokes on an image with different colors (8 choices) and pen sizes (4 choices); the dataset consists of 436 dermoscopic images in standard JPEG format with a resolution of 4000×2664 or 3000×4000 pixels. Large part of design and technology is in common with the original annotation tool and the use of the .NET Framework (ASP.NET with C# the server side, XHTML with Javascript the client side) allows to neglect the problem of the *input mode*: since the client module reaches the server one by a standard web

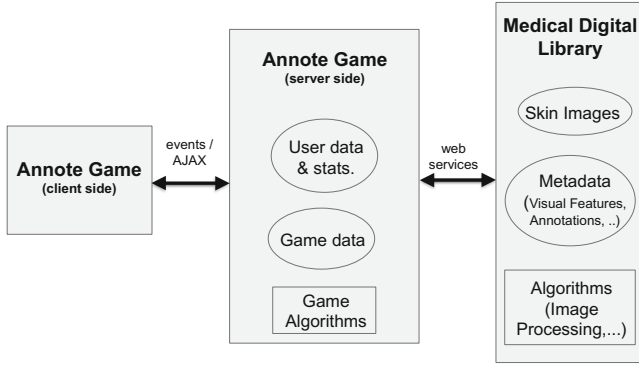


Fig. 2. The system architecture

connection it is possible to interact with the *Surface Pen* if the client runs on a Surface device or with mouse/touch otherwise.

In Fig. 2 there is the system architecture: a *Digital Library* manages the image collection, the ground-truth annotations and the algorithms for visual meta-data as in [4, 6, 10]; it interfaces with the game server using *web services* that allow communication between heterogeneous technologies; the game client permits the interaction with the user interface exploiting dynamic events and AJAX requests.

The *Annote server* manages data about the registered students with their *points* and *achievements*; the *game data* repository stores the settings of each game task while the *game algorithms* module performs evaluations (for example comparing an ‘official ground-truth’ annotations with that made by the player).

All gaming interactions and stroke collections are managed locally on the client and the data exchange with the server is limited to the original skin image, the drawn strokes and all the textual data (points, time, task requests, settings): this permits to minimize the network resources and to separate the management of data and algorithms from the module used by students to learn and to experiment on real medical data (that results protected for the privacy). In this prototype the user interface is minimal but functional and it is divided into three main sections (Fig. 3):

- upper section: shows the player profile (photo, nickname, points, badges, ...)
- middle section: tools to manage the image (Adapt, Center, Zoom) and to change the interaction mode (Erase or Draw) or the stroke features (color/width selection, annotation deleting)
- lower-left section: shows the main skin image with the superimposed strokes
- lower-right section: represents the interactive section of the game, in fact shows points and time (for time challenges), the setting of the task and the interactive messages occurred during the gaming session; moreover there are three buttons to commit, reload or change the game session.



Fig. 3. The user interface of the Annote serious game

At the beginning of a game session, an XML configuration file is sent by the server to set up the user interface (for example to enable/disable buttons and widgets, instantiate the timer, load the image or draw default strokes). To manage game dynamics like *power-up* or the *increasing difficulty*, the game client implements a simple *event manager* that sends asynchronous messages to the server which in turn raises appropriate counter-events: for example, the end of the time involves a game failure, a *zoom event* will enable/disable the corresponding buttons, a *time X event* will increase/decrease the timer by X seconds, a *speed X event* will increase/decrease the mouse/pen speed by an amount of X. When a student commits his work, the server algorithms will compute the corresponding rewards and will update the total points of the player in the leaderboard rank of the registered gamers.

4 Conclusions and Future Work

In this paper, we exploited the interesting aspect of a pedagogical use of a digital library: help academic physicians in the delicate task of teaching, by transmitting a tacit knowledge hard to express with standard educational modes.

After the prototype presented here, it is necessary to design and perform evaluation studies on the field, involving directly large amount of medical students to obtain qualitative and quantitative data and compare the progressions between the ‘standard’ learning and the ‘gamified’ one. A usable *editor* can help to better customize the contents as well as services that address to academic resources when students have doubts; moreover sound and graphic effects with some forms of cooperative or multiplayer features can expand and improve the

gaming experience. Finally, the expansion of this gamification process to other medical specializations that involve annotation protocols may well deserve further insights.

References

1. Ardito, C., Buono, P., Costabile, M.F., Lanzilotti, R., Pederson, T.: Re-experiencing history in archaeological parks by playing a mobile augmented reality game. In: Meersman, R., Tari, Z., Herrero, P. (eds.) OTM 2007. LNCS, vol. 4805, pp. 357–366. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76888-3_58
2. Balducci, F., Borghi, G.: An annotation tool for a digital library system of epidermal data. In: Grana, C., Baraldi, L. (eds.) IRCDL 2017. CCIS, vol. 733, pp. 173–186. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68130-6_14
3. Balducci, F., Grana, C.: Affective classification of gaming activities coming from RPG gaming sessions. In: Tian, F., Gatzidis, C., El Rhalibi, A., Tang, W., Charles, F. (eds.) Edutainment 2017. LNCS, vol. 10345, pp. 93–100. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65849-0_11
4. Balducci, F., Grana, C.: Pixel classification methods to detect skin lesions on dermoscopic medical images. In: Battiato, S., Gallo, G., Schettini, R., Stanco, F. (eds.) ICIAP 2017. LNCS, vol. 10485, pp. 444–455. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68548-9_41
5. Baraldi, L., Grana, C., Cucchiara, R.: Recognizing and presenting the storytelling video structure with deep multimodal networks. *IEEE TMM* **19**(5), 955–968 (2017)
6. Bolelli, F., Cancilla, M., Grana, C.: Two more strategies to speed up connected components labeling algorithms. In: Battiato, S., Gallo, G., Schettini, R., Stanco, F. (eds.) ICIAP 2017. LNCS, vol. 10485, pp. 48–58. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68548-9_5
7. Callaghan, M.J., McShane, N., Eguíluz, A.G., Teillès, T., Raspail, P.: Practical application of the learning mechanics-game mechanics (LM-GM) framework for serious games analysis in engineering education. In: 2016 13th International Conference on Remote Engineering and Virtual Instrumentation (REV), pp. 391–395, February 2016
8. Chen, J.: Flow in games (and everything else). *Commun. ACM* **50**(4), 31–34 (2007)
9. Coltel, O., Grandi, X., Tosca, R., Latorre, P., Sánchez, J.S., Lizán, L.V., Ros-Bernal, F., Martínez-Cadenas, C.: Designing serious games for learning support in medicine studies: a specific method to elicit and formalize requirements. In: 2014 IEEE Frontiers in Education Conference (FIE), pp. 1–4. IEEE (2014)
10. Corbelli, A., Baraldi, L., Balducci, F., Grana, C., Cucchiara, R.: Layout analysis and content classification in digitized books. In: Agosti, M., Bertini, M., Ferilli, S., Marinai, S., Orío, N. (eds.) IRCDL 2016. CCIS, vol. 701, pp. 153–165. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56300-8_14
11. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Visual saliency for image captioning in new multimedia services. In: IEEE International Conference on Multimedia and Expo Workshops (2017)
12. Csikszentmihályi, M.: *Beyond Boredom and Anxiety*. Jossey-Bass, San Francisco (2000)
13. Cucchiara, R., Grana, C., Prati, A.: Semantic transcoding for live video server. In: Proceedings of the 10th ACM International Conference on Multimedia, MULTIMEDIA 2002 (2002)

14. Deponti, D., Maggiorini, D., Palazzi, C.E.: Smartphone's psychiatric serious game. In: IEEE International Conference on Serious Games and Applications for Health (2011)
15. Hamari, J., Koivisto, J., Sarsa, H.: Does gamification work?-A literature review of empirical studies on gamification. In: Hawaii International Conference on System Sciences (2014)
16. Huotari, K., Hamari, J.: Defining gamification-a service marketing perspective. *System* **1**(2), 3–4 (2012)
17. Maggiorini, D., Quadri, C., Ripamonti, L.A.: Opportunistic mobile games using public transportation systems: a deployability study. *Multimed. Syst.* **20**(5), 545–562 (2014)
18. Mäyrä, F.: *An Introduction to Game Studies*. SAGE Publications, London (2008)
19. Nakamura, J., Csíkszentmihályi, M.: The concept of flow. In: *Handbook of Positive Psychology*, pp. 89–105 (2002)
20. Norman, D.A.: Affordance, conventions, and design. *Interactions* **6**(3), 38–43 (1999)
21. Shneiderman, B.: 1.1 direct manipulation: a step beyond programming languages. *Sparks Innov. Hum. Comput. Interact.* **17**, 1993 (1993)
22. Ushaw, G., Eyre, J., Morgan, G.: A paradigm for the development of serious games for health as benefit delivery systems. In: 2017 IEEE 5th International Conference on Serious Games and Applications for Health (SeGAH), pp. 1–8. IEEE (2017)

Term-Based Approach for Linking Digital News Stories

Muzammil Khan¹(✉), Arif Ur Rahman², and Muhammad Daud Awan¹

¹ Department of Computer Science, Preston University, Islamabad, Pakistan
muzammilkhan86@gmail.com, drdaudawan@preston.edu.pk

² Department of Computer Science, Bahria University, Islamabad, Pakistan
badwanpk@gmail.com

Abstract. The World Wide Web has become a platform for news publication in the past few years. Many television channels, magazines and newspapers have started publishing digital versions of the news stories online. It is observed that recommendation systems can automatically process lengthy articles and identify similar articles to readers based on a predefined criteria i.e. collaborative filtering, content-based filtering approach. The paper presents a content-based similarity measure for linking digital news stories published in various newspapers during the preservation process. The study compares similarity of news articles based on human judgment with a similarity value computed automatically using common ratio measure for stories. The results are generalized by defining a threshold value based on multiple experimental results using the proposed approach.

Keywords: Linking news stories · Similarity measures
Text processing

1 Introduction

The advanced technologies and proliferation of the Internet attract news readers to read online news from multiple sources and get the desired information. It is not humanly possible to browse through a huge information space for related information items. The amount of news article releases has grown rapidly and for an individual it is cumbersome to browse through all online sources for relevant news articles. Information retrieval techniques help in searching through the vast information spaces up to some extent and recommendation systems have emerged to respond the challenge by providing users the information which matches their needs either by their preferences or by content similarity among the news items. Each online news provider tries to handle their news articles and use some mechanisms to recommend similar news to the news readers.

The news generation in the digital environment is no longer a periodic process with a fixed single output like printed newspaper. The news are instantly generated and updated online in a continuous fashion. However, because of different

reasons like the short lifespan of digital information and speed of generation of information, it has become vital to preserve digital news for the long-term.

Digital preservation includes various actions to ensure that digital information remains accessible and usable, as long as they are considered important [3]. Libraries and archives preserve newspapers by carefully digitizing collections as newspapers are a good source of knowing history. Many approaches have been developed to preserve digital information for the long term [5, 15]. The lifespan of news stories published online vary from one newspaper to another, i.e. from one day to a month. Though, a newspaper may be backed-up and archived by the news publisher or national archives, in the future it will be difficult to access particular information published in various newspapers about the same news story. The issues become even more complicated if a story is to be tracked through an archive of many newspapers, which require different access technologies.

The Digital News Story Extractor (DNSE) is a tool developed to facilitate the extraction of news stories from the online newspapers and migrate to a normalized format for preservation using Digital News Stories Preservation (DNSP) framework [8]. The normalized format also includes a step to add metadata in the Digital News Stories Archive (DNSA) for future use [9]. To facilitate the accessibility of news articles preserved from multiple sources, some mechanisms needs to be adopted for linking the archived digital news articles.

The study proposes an effective term-based approach for linking digital news articles in DNSA. The approach is empirically analyzed, and the results of the proposed approach are compared to get conclusive arguments.

2 Background

Enormous information is available on the web for users, including a variety of products and options in the form of books, restaurants, hotels, research articles, movies, news articles, etc. Recommender systems help users to focus down information to manageable sets. Broadly, two approaches have been devised, Collaborative Filtering approach, based on similar users having similar interest or same demographics and Content-based approach, which is based on features of the item to be recommended [2, 13].

It is observed that the news articles available can be very huge and recommendation systems can help to recommend relevant news to news readers by filtering news articles based on predefined criteria. There are two approaches that can be used to link news articles together either by collaborative filtering or content based approach. Collaborative filtering method presents many challenges as it relies on the similarity in opinions and demographics of the users [6]. It becomes more complicated with dynamic nature of the users and news articles themselves. Users prefer to find recent news in online news environment, which is difficult to learn user's preferences that lead an accurate model based on the items they previously read [1, 7, 11]. User preferences and interest changes over time, depends on the current events and popularity of the news articles themselves [12]. Typically, users are not willing to click, to recommend news articles

during browsing news on the specific topic [16]. Content-based approach recommends new items to the user based upon the similarity value being computed between the descriptions or features of items selected previously. Content based approach can run through its own problems like determine similarity between news articles that represents different topics and the way user's choice effect by some potentially hidden factors [10]. All these studies focused on run time similarity between recent articles. The subsequent section describes the linkage of preserved news articles in the DNSA.

2.1 Content-Based Similarity Approach

Collaborative filtering approach suffers through a number of challenges because of many reasons. In contrast, content-based filtering approach can be used to adopt features of the item, i.e. news article to be recommended or link articles together by some predefined criteria of similarity among the news stories by extracting features from news articles to avoid third party dependency, i.e. user feedback, etc. Almost all the studies about recommending news articles focuses on run time similarity and recommendation because these platforms have dealt with few numbers of recently published articles. The subsequent section describes the linkage of preserved news articles in the DNSA.

3 Linking Digital News Stories in DNSA

The Digital News Stories Archive (DNSA) has passed more than six months of the trial period. The number of extraction has passed greater than 100 times at different intervals in which initially three locals leading English online newspapers (Dawn News, The Tribune and The News) being considered for preservation and the number increased to ten, that include seven local English online newspapers and three local news television networks, which provide online English news to the news readers. The archive (created locally) currently preserving more than one thousand news after removing duplicate URLs and news in each extraction.

The news readers read about an event or an issue from various sources in order to get a broader perspective and diverse viewpoints that help to better understand the world around. Moreover, consulting various sources helps in authenticating the information by comparing similar news from multiple news sources. The DNSA has news articles from multiple sources, needs to create a mechanism that helps the reader to read a set of relevant news stories about an event or issue. The DNSA needs an efficient mechanism to link the digital stories and recommend to the readers. This linkage will lead the reader to browse through the huge collection easily. Without a suitable and efficient linking mechanism for relevant news, the newspaper or online newspaper archive is nothing more than a data collection.

A link can be created by two means, namely vertical linkage and horizontal linkage as presented in Fig. 1.

1. **Vertical Linkage:** The link is created between the news stories which represent the “same story” regardless of the time frame based on all the news preserved in the DNSA. Same story can be interpreted as the follow-up news about an issue from the same source or from different sources. For example, news stories about the court proceedings of a particular case. Same story linkage leads to follow-up similarity between news for more days since the start of proceedings.
2. **Horizontal Linkage:** The link is created between the news stories which represent the “same news” based on all the news preserved in the DNSA in a specific time period i.e. day, week, month. Same news can be interpreted as the similar news reported about an issue from the same source or from different sources. For example, Court response to the arguments about a particular case. Same news linkage makes the identification of news about the same topic easy.

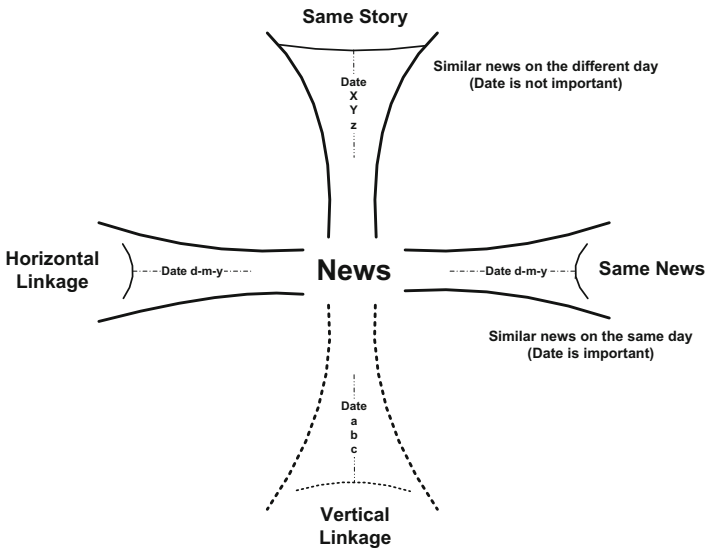


Fig. 1. Linking digital news stories in DNSA

In this study, the focus is to create links between relevant stories based on similar terms used in the news articles, comprehensively discussed subsequently.

The news articles are linked in the DNSA during the preservation process in the form of implicit metadata to easily relate news stories based on horizontal linkage. It is observed that two news stories about an event may contain similar terms and published online on the same day, which show the importance of terms used and encourage utilizing these terms for linking digital news stories during preservation in the Digital News Stories Archives.

To speed up the linking process of digital news stories for preservation, a tiered approach is adopted as presented in Fig. 2.

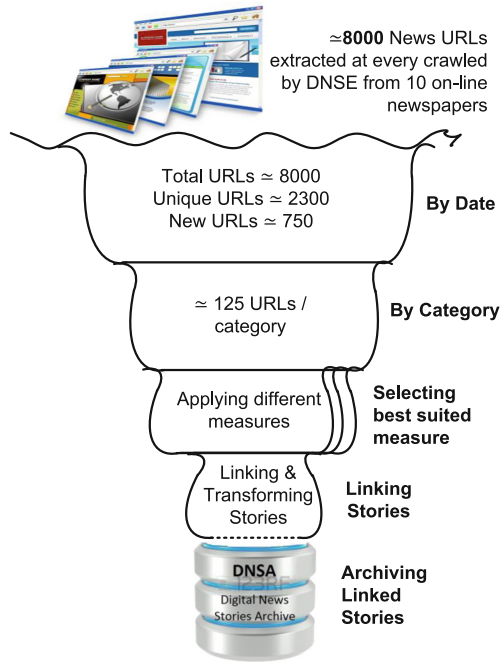


Fig. 2. Layered approach to reduce workload during linking digital stories for preservation

The main purpose of the tiered approach is to reduce the linking workload and improve performance. Currently, the DNSA archives 10 online newspapers and more than 8000 URLs are extracted at every crawl by DNSE. In the first tier, the news URLs are filtered by date, so that removing duplicate URLs and news already archived from previous crawls. In the second tier, the news is filtered by categorizing the news articles into six known categories which reduce the news article collection sufficiently for linking. Third tier can be divided into multiple tiers based on different measures developed for linking digital news stories for preservation. The selection of measure depends upon the performance of that measure in a specific category, e.g. a measure can perform better on a specific category, such as “Opinion”. In the last tier, the news stories are linked and transformed to the digital news stories archive.

4 Term-Based Approach

To find out the similarity between news in the DNSA, needs to process the terms used in the news articles. The news article contains the different type of terms,

e.g. nouns, verbs, adverbs, etc. In academic journals, the nouns are considered to be the main key phrases [14] but other terms like verbs and adverbs, etc., also play a vital role in representing the news articles [4]. Therefore, introduced a Common Ratio Measure for Stories (CRMS) based on the similar terms used in the English news articles except stop words for linking digital news stories during preservation. The CRMS Algorithm 1 pseudo-code is given as follows;

Algorithm 1. CRMS Algorithm

- 1: **News article pre-processing**
 - 2: Filtering non-news contents and extracts the news article from the news webpage
 - 3: **Compute Term Frequencies**
 - 4: **repeat**
 - 5: Tokenize news articles using StanfordCoreNLP
 - 6: Remove stop words
 - 7: Calculate term frequencies of each term in the news articles
 - 8: **until** Both the news articles are processed
 - 9: **Compute CT, UT and TT**
 - 10: Compute CT (Common Terms) Count
 - 11: Select all common terms in both the news articles with frequencies
 - 12: $CT = (tf_1 + tf_2)W_1 + (tf_1 + tf_2)W_2 + \dots + (tf_1 + tf_2)W_n$
 - 13: $CT = \sum_{i=1}^n (tf_1 + tf_2)W_i$
 - 14: ▷ Where, W_i is the common term or word in both the selected news articles, tf_1 term frequency of word W in one news, tf_2 is term frequency of word W in second news article and n is the total number of common terms in both the news.
 - 15: Compute UT (Uncommon Terms) Count
 - 16: Select all uncommon terms in both the news articles with frequencies
 - 17: $UT = (tf_1 \vee tf_2)W_1 + (tf_1 \vee tf_2)W_2 + \dots + (tf_1 \vee tf_2)W_m$
 - 18: $UT = \sum_{j=1}^m (tf_1 \vee tf_2)W_j$
 - 19: ▷ Where m is the total number of uncommon terms in both the news
 - 20: Compute TT (Total Terms) Count
 - 21: $TT = CT + UT = \sum_{i=1}^n (tf_1 + tf_2)W_i + \sum_{j=1}^m (tf_1 \vee tf_2)W_j$
 - 22: ▷ The total terms in both the news articles
 - 23: **Compute Common Ratios**
 - 24: Four common ratios can be used as similarity measure between news articles
 - 25: $CRMS = CT/TT$
 - 26: $CRMS = CT/UT$
 - 27: $CRMS = UT-CT$
 - 28: $CRMS = UT/TT$
-

CT/TT : The value varies between 1 ($CT = TT$) and 0 ($CT = 0$ i.e. $UT = TT$). 1 means the news are exact copies of each other and 0 represents that the stories are completely different. Closer the value CT/TT to 1 leads to more similarity and

closer the value to 0 decrease similarity. The interpretation of measure UT/TT is exactly opposite to the measure CT/TT . The value CT needs to normalize by the value TT because the high value of CT only does not show the accurate similarity.

For example, if CT of two news A and B is 150 and CT of news A and C is 100 do not show that the A and B is more similar than A and C until normalized. If TT count for A & B is 400 and for A & C its 200 then by CT/TT are 0.375 and 0.5 respectively, shows that news articles A & C is more similar than A & B.

CT/UT : The value varies between 0 ($CT = 0$) and TT , $UT > 0$, greater the value similar the news articles and vice versa.

$UT - CT$: The minimum possible value is $-TT$ and maximum value is $+TT$. Minimum the value leads to high similarity of the news articles and maximum the value leads to dissimilarity.

5 Evaluating CRMS

The size of the DNSA is growing very fast as on average seven hundred new articles are add every day. The number may raise even further when more news sources are added. The size of the DNSA makes it an ideal choice for evaluating the CRMS. Various news sets were created for evaluating the significance of CRMS.

5.1 News Sets

The following four sets were defined which were created multiple times for evaluating the CRMS.

- **Set 1:** A set containing three news articles - two similar and one dissimilar. Each article in this set is manually chosen from a different newspaper.
- **Set 2:** A set containing ten news articles - three closely related, one partially similar and six dissimilar. The articles in this set are manually chosen from three different newspapers.
- **Set 3:** A set containing thirty news articles - the news articles are categorized based on six topics. Table 1 presents the details of the stories in each category. The articles in this set are manually chosen from nine different newspapers.
- **Set 4:** A set containing 215 news articles - the articles are grouped into two categories. First, contains 52 articles related to sports which were automatically extracted from the DNSA using the category information. Second contains 163 automatically extracted news articles related to other topics including politics, business, law and order situation, crimes and cases in courts. The articles in this set are chosen randomly from three different newspapers.

Table 1. News articles distribution in 30 news articles dataset

S.No	Topic	No of news
Topic 1	Disruptive passenger in PIA at Heathrow London	6
Topic 2	Trump travel ban	5
Topic 3	CPEC	5
Topic 4	Nurses protest in Karachi	4
Topic 5	Earthquake in Baluchistan	5
Topic 6	LoC ceasefire violation	5

5.2 Evaluation Methods

Two evaluation methods, namely system centric (automatic) and user centric (human judgment) were used for assessing the accuracy and effectiveness of the designed similarity measure for linking news stories for preservation.

1. Pre-experiment: Human Judgment of Similarity

In this section the similarity level by human judgment is defined. The similarity between the news articles has defined empirically by two means, i.e. online news readers and by an expert. The online news readers included graduate and undergraduate students as well as faculty members who were randomly selected to rank the similarity among news articles based on human judgment. An individual is selected as an expert who knows comprehensively about the DNSE and the process of digital news preservation in DNSA.

Three sets of news articles were created and used to evaluate the proposed approach. Each set contained ten news articles selected from different online news publishers. The collection contained news articles that were selected by reading headings from the same genre i.e. sports, opinions, entertainment. The detail of each news set is presented in Sect. 5.1.

To define similarity among news articles, the participants were asked to define similarity of one news article with the rest of the news articles in the collection using five-point Likert scale and a representation of numerical measure from 1 to 10. The basic reason to consider the numerical scale to easily differentiate similar or dissimilar news i.e. if news B & C are marked similar to that of news A in Likert scale; the same may be marked as 10 and 9 respectively at the numeric scale that shows the news B is more similar to that of news C. The following Table 2 show the numerical equivalent values of likert scale.

Table 2. Numeric scale to corresponding likert scale

Likert scale	Similar	Partially similar	Unsure	Partially dissimilar	Dissimilar
Numeric scale	10–9	8–7	6–5	4–3	2–1

2. Automatic Evaluation

Precision and recall were calculated and compared with the proposed similarity measure i.e. CRMS. *Precision* is the ratio of the number of relevant items retrieved to the total number of irrelevant and relevant items retrieved. *Recall* is the ratio of the number of relevant items retrieved to the total number of relevant items in a collection. Let A is No of relevant items retrieved, B is No of irrelevant items retrieved and C is No of relevant items not retrieved, the set-based measures are given in Table 3.

Table 3. Set-based measures

	Relevant	Not relevant
Retrieved	A	B
Not retrieved	C	D

$$\text{Precision} = A/A + B(\text{Rel}\&\text{Ret}/\text{Retrieved})$$

It is usually expressed in percentage i.e.

$$\text{Precision} = (A/A + B) * 100$$

$$\text{Recall} = A/A + C(\text{Rel}\&\text{Ret}/\text{Relevant})$$

It is usually expressed in percentage i.e.

$$\text{Recall} = (A/A + C) * 100$$

5.3 Results

The proposed term based approach i.e. CRMS, for similarity between news articles is analyzed on different sets of news articles. The overview of datasets is summarized in the Table 4.

Table 4. Overview: Datasets of news articles used for evaluation

S.No	News articles			Similarity observed			
	No of news articles/set	No of sets	No of newspapers	During selection	By news reader	By expert	Proposed measures
1	3	3	3	Yes	No	No	Yes
2	10	3	3	Yes	Yes	Yes	Yes
3	30	1	9	Yes	No	No	Yes
4	215	1	3	No	No	No	Yes

For each set of news articles, similarity is computed and measured by two means, i.e. empirically (User based) and Automatic (using CRMS algorithm). The Tables 5, 6 and 7 shows the summary of the values computed for evaluation and the similarity between news articles are compared.

An experiment was performed using a set of articles collected following the definition of set three (defined in Sect. 5.1). The performance of CRMS for the set is presented in the Table 8.

Table 5. Similarity comparison (Likert scale) with CRMSs for Set 1

News1	News Readers		Expert		Common Ratio Measure for Stories (CRMS)							
	News	Mean	News	Value	CT/TT	Value	UT/TT	Value	CT/UT	Value	UT-CT	Value
ns1	ns3	4.7	ns3	5	ns3	0.542	ns3	0.458	ns3	1.181	ns3	-29
ns1	ns8	4.3	ns8	5	ns8	0.409	ns8	0.591	ns8	0.691	ns8	81
ns1	ns5	3.6	ns5	4	ns7	0.333	ns7	0.667	ns7	0.498	ns7	149
ns1	ns10	2.9	ns10	4	ns10	0.262	ns10	0.738	ns10	0.355	ns10	151
ns1	ns7	2.7	ns7	4	ns5	0.222	ns5	0.778	ns5	0.285	ns9	166
ns1	ns9	2.6	ns9	4	ns9	0.165	ns9	0.835	ns9	0.198	ns4	177
ns1	ns4	2.4	ns4	4	ns6	0.161	ns6	0.839	ns6	0.192	ns5	181
ns1	ns6	1.4	ns6	1	ns4	0.147	ns4	0.853	ns4	0.173	ns2	201
ns1	ns2	1.3	ns2	1	ns2	0.135	ns2	0.865	ns2	0.155	ns6	282

Table 6. Similarity comparison (Likert scale) with CRMSs for Set 2

News1	News Readers		Expert		Common Ratio Measure for Stories (CRMS)							
	News	Mean	News	Value	CT/TT	Value	UT/TT	Value	CT/UT	Value	UT-CT	Value
ns1	ns5	4.6	ns5	5	ns5	0.281	ns5	0.719	ns5	0.39	ns5	97
ns1	ns9	4.3	ns9	4	ns2	0.158	ns2	0.842	ns2	0.187	ns4	116
ns1	ns10	3.4	ns10	4	ns8	0.154	ns8	0.846	ns8	0.182	ns3	119
ns1	ns7	1.7	ns7	1	ns7	0.148	ns7	0.852	ns7	0.174	ns2	152
ns1	ns6	1.6	ns6	1	ns10	0.141	ns10	0.859	ns10	0.164	ns6	152
ns1	ns2	1.4	ns2	1	ns9	0.128	ns9	0.872	ns9	0.147	ns8	157
ns1	ns3	1.4	ns3	1	ns6	0.116	ns6	0.884	ns6	0.131	ns9	215
ns1	ns8	1.4	ns8	1	ns3	0.066	ns3	0.934	ns3	0.07	ns7	223
ns1	ns4	1.2	ns4	1	ns4	0.032	ns4	0.968	ns4	0.033	ns10	229

Table 7. Similarity comparison (Likert scale) with CRMSs for Set 3

News1	News Readers		Expert		Common Ratio Measure for Stories (CRMS)							
	News	Mean	News	Value	CT/TT	Value	UT/TT	Value	CT/UT	Value	UT-CT	Value
ns1	ns3	4.4	ns3	5	ns3	0.688	ns3	0.312	ns3	2.203	ns3	-273
ns1	ns7	4.1	ns2	5	ns7	0.41	ns7	0.59	ns7	0.694	ns7	89
ns1	ns2	3.7	ns5	5	ns5	0.407	ns5	0.593	ns5	0.686	ns5	91
ns1	ns4	3.7	ns7	4	ns2	0.405	ns2	0.595	ns2	0.68	ns4	118
ns1	ns5	3.6	ns4	4	ns4	0.372	ns4	0.628	ns4	0.593	ns2	129
ns1	ns8	3.5	ns8	4	ns8	0.308	ns8	0.692	ns8	0.445	ns8	177
ns1	ns9	3.4	ns9	4	ns6	0.256	ns6	0.744	ns6	0.345	ns6	251
ns1	ns10	2.9	ns10	4	ns10	0.256	ns10	0.744	ns10	0.345	ns9	278
ns1	ns6	2.4	ns6	4	ns9	0.175	ns9	0.825	ns9	0.212	ns10	287

The results of the evaluation show that the CRMS gives reliable results. Therefore, it is useful to use the CRMS for linking digital news stories. An archive containing preserved news stories which are linked using the proposed measure i.e. CRMS will support accessibility of related news articles.

To measure the precision and recall of the proposed similarity measure i.e. CRMS, the experiment is performed on a set of articles collected following the definition of Set three (defined in Sect. 5.1). The similarity is observed by an expert (manually) during the selection of news articles for the experiments as presented in Table 4. The performance of CRMS for the dataset is presented in the Table 8.

Table 8. Precision and recall for CRMS

S.No	Topic	Precision	Recall
Topic 1	Disruptive passenger in PIA at Heathrow London	100%	100%
Topic 2	Trump travel ban	80%	100%
Topic 3	CPEC	60%	75%
Topic 4	Nurses protest in Karachi	60%	100%
Topic 5	Earthquake in Baluchistan	80%	100%
Topic 6	LoC ceasefire violation	80%	100%

The results of the evaluation show that the CRMS gives reliable results. Therefore, it is useful to use the CRMS for linking digital news stories. An archive containing preserved news stories which are linked will support accessibility of related news articles. The digital news stories normalized, linked and preserved ensure accessibility of related news articles.

6 Conclusions and Future Work


The proposed measure i.e. CRMS is simple to calculate as it involves just a few steps. However, the proposed algorithm may not produce accurate results when used for two articles having very different lengths i.e. compare a two sentence story with a full page story. However, the findings show that the similarity results based on the CRMS are able to capture the reality when used for articles not very different in length.

Currently, work is going on to extend the linkage of stories to news published in Urdu language. Moreover, work is in progress to develop tools for exploiting the linkage created among stories during the preservation process for search and retrieval tasks.

References

1. Agarwal, D., Chen, B.-C., Elango, P., Wang, X.: Personalized click shaping through lagrangian duality for online recommendation. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 485–494. ACM (2012)
2. Athalye, S.: Recommendation system for news reader (2013)
3. Burda, D., Teuteberg, F.: Sustaining accessibility of information through digital preservation: a literature review. *J. Inf. Sci.* **39**(4), 442–458 (2013)
4. Chun, D.: On indexing of key words. *Acta Editologica* **16**(2), 105–106 (2004)
5. da Silva, J.R., Ribeiro, C., Lopes, J.C.: A data curation experiment at U. Porto using DSpace (2011)
6. Doychev, D., Lawlor, A., Rafter, R., Smyth, B.: An analysis of recommender algorithms for online news. In: CLEF (Working Notes), pp. 825–836. Citeseer (2014)
7. Fortuna, B., Fortuna, C., Mladeníć, D.: Real-time news recommender system. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010. LNCS (LNAI), vol. 6323, pp. 583–586. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15939-8_38
8. Khan, M., Ur Rahman, A.: Digital news story preservation framework. In: Proceedings of Digital Libraries: Providing Quality Information: 17th International Conference on Asia-Pacific Digital Libraries, ICADL 2015, Seoul, Korea, 9–12 December 2015, vol. 9469, p. 350. Springer (2015). <https://doi.org/10.1007/978-3-319-27974-9>
9. Khan, M., Ur Rahman, A., Daud Awan, M., Alam, S.M.: Normalizing digital news-stories for preservation. In: 2016 Eleventh International Conference on Digital Information Management (ICDIM), pp. 85–90. IEEE (2016)
10. Li, L., Li, T.: News recommendation via hypergraph learning: encapsulation of user behavior and news content. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 305–314. ACM (2013)
11. Li, L., Wang, D.-D., Zhu, S.-Z., Li, T.: Personalized news recommendation: a review and an experimental investigation. *J. Comput. Sci. Technol.* **26**(5), 754–766 (2011)
12. Li, L., Zheng, L., Yang, F., Li, T.: Modeling and broadening temporal user interest in personalized news recommendation. *Expert Syst. Appl.* **41**(7), 3168–3177 (2014)
13. Melville, P., Sindhvani, V.: Recommender systems. In: Sammut, C., Webb, G.I. (eds.) *Encyclopedia of Machine Learning*, pp. 829–838. Springer, Heidelberg (2011). https://doi.org/10.1007/978-0-387-30164-8_705
14. Pu, P., Chen, L., Hu, R.: A user-centric evaluation framework for recommender systems. In: Proceedings of the Fifth ACM Conference on Recommender Systems, pp. 157–164. ACM (2011)
15. Ur Rahman, A.: Data warehouses in the path from databases to archives. Ph.D. thesis, Faculty of Engineering, University of Porto, July 2013
16. Said, A., Bellogín, A., Lin, J., de Vries, A.: Do recommendations matter?: news recommendation in real life. In: Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 237–240. ACM (2014)

A Graphic Matching Process for Searching and Retrieving Information in Digital Libraries of Manuscripts

Nicola Barbuti¹, Tommaso Caldarola², and Stefano Ferilli³

¹ Department of Humanities (DISUM), University of Bari Aldo Moro, Bari, Italy
nicola.barbuti@uniba.it

² D.A.BI.MUS. Ltd., Spin Off of University of Bari Aldo Moro, Bari, Italy
t.caldarola@dabimus.com

³ Department of Computer Science (DIB), University of Bari Aldo Moro, Bari, Italy
stefano.ferilli@uniba.it

Abstract. This paper outlines ICRPad, a pattern recognition system based on a graphic matching algorithm, which works on images by shape contour recognition, without requiring any segmentation process. The algorithm starts the process from a region of interest (ROI) selected in the image, using it as a shape model and looking for similar patterns in one or many target images. The process was developed and tested with the aim of proposing a new approach for searching and retrieving information in digital libraries. This approach is based on the application of data science, the fourth paradigm of knowledge development in the scientific field, that is at the basis of science informatics, to studies in data humanities. Following this approach, the algorithm is applied to find new research hypotheses through the discovery of patterns directly inferred from large digital libraries.

Keywords: Graphic pattern · Pattern recognition · Digital libraries · Manuscripts
Graphic matching algorithm

1 Introduction

Historically, the development of knowledge in the scientific field has been carried out following two paradigms, the theoretical one and the experimental one. In the last two decades two additional paradigms have been established: the computer-based simulation one (*computational simulation* – Ken Wilson, Nobel prize in physics, 1982), also called the ‘third paradigm’, and the one based on data-driven scientific discovery (*data intensive scientific discovery* – Gordon Bell, 2012), also known as the ‘fourth paradigm’. While the third paradigm gave rise to *computational sciences* (e.g., computational biology), the fourth one is at the basis of *science informatics* (e.g., bioinformatics). The latter has gained acceptance thanks to the increasing availability of huge amounts of data that permit an *in silico* approach to knowledge generation.

To apply the same change of perspective in the humanities, one must start by observing that in the last decades significant effort has been spent in generating large humanistic databases that can be accessed online. Some examples are:

- *Thesaurus Linguae Graecae*, that collects Greek literature since Homer (VIII sec. BC) to the fall of Byzantium (1453 AD) [<http://stephanus.tlg.uci.edu/>];
- *Integrated Archaeological Database* (IADB), that addresses the data management requirements throughout the lifespan of archaeological excavation projects [<http://www.iadb.org.uk/>];
- *World Digital Library* (WDL), that collects digitized versions of rare books, maps, manuscripts, and photographs [<https://www.wdl.org/en/>];
- *Musisque Deoque*, a digital archive of Latin poetry [<http://www.mqdq.it/public/>];
- *Trismegistos*, a database concerning writings on papyrus [<http://www.trismegistos.org/>].

All these databases propose querying mechanisms of different levels of complexity, that mostly provide support to scholars in their specific searches (e.g., retrieving all poems written using a given prosody). However, this requires that scholars have previously set up accurate hypotheses that they want to confirm by searching through the digital archives. An opposite methodological approach is proposed by the fourth paradigm: Algorithms are applied to find new working hypotheses through the discovery of ‘*patterns*’ directly inferred from large databases. For instance, groups (‘clusters’) of ‘similar’ poems might be identified in digital libraries or archaeological collections, or in literary corpora, and suggest new hypotheses on which an enquiry using traditional approaches can be started.

This paper presents an application of the fourth paradigm for allowing scholars to query and search large digital libraries of manuscripts using ICRPad, a patented digital recognition system. It embeds a graphic matching algorithm which works on images by shape contour recognition, without requiring any segmentation process of the image content [1]. The system has been tested on a dataset of manuscripts, and the results suggest the viability of a new approach to studies in data humanities.

2 Related Works

Contemporary digital databases commonly use recognition systems to convert digital images into machine-encoded texts, but to date no system works efficiently on images of manuscripts. Many research projects have been devoted to solve this problem by creating OCR or pattern recognition systems for digital cultural heritage. Nevertheless, these technologies have not yet shown sufficient effectiveness and functional efficiency so as to provide an immediately accessible indexing of the image content.

Although much research has been carried out, digital recognition has been successful on small databases and highly constrained domains only. There is not yet any valid system for querying, recognizing and searching large databases of historical manuscripts. Research is mainly based on two different approaches, either segmental or holistic. The segmental approach is undoubtedly the most used in a number of recent systems, especially in prototypes structured according to Hidden Markov Models (HMM) [2–7]. The holistic approach was preferred in some recent experiments, with results of great interest as regards the percentage of recognized content, but not significant as regards the number of processed images [8].

In detail, most research is based on:

- segmentation – shape models are created from segmented regions (portions of graphemes, graphemes, words, etc.) and then classified by reference to thesauri that support text matching, an both laborious and potentially very time-consuming operation;
- adaptation of existing processes (word spotting, HMMs, etc.) – overcomes text segmentation, but requires data extraction and matching processes that can detect and recover functions referring to statistical criteria; moreover, most of these prototypes run mainly on digital images of printed documents, but their effectiveness and usability on manuscripts are unknown;
- matching each word found in digital images either to the corresponding electronic text previously transcribed by an operator manually, or to reference thesauri of selected words preliminarily structured (again, manually) – this approach seriously limits the possibility of electronically recognizing a large quantity of historical texts, because it requires a long and complex preliminary manual work [9–20].

Further limits of these prototypes are the following:

- they have been tested on very small quantities of images, so there is no proof of their applicability to large digital databases;
- the above methodologies provide preliminary quite complex and time consuming human work, without noteworthy results, scarcely useful: indeed, excessive manual work greatly limits the possibility of indexing large quantities of images, requiring a great deal of human and financial resources;
- the research proposed purely theoretical models without any certainty about their effective ability of working on digital databases;
- no prototype really uses automatic or semi-automatic recognition: all prototypes need a preliminary planning of complex algorithms to extract information and create the models by which performing the matching with digital images, but their output is nearly always incomplete and unsatisfactory.

Unlike the systems and prototypes described above, the ICRPad graphic matching algorithm proposes a different way for creating the shape model, based on contour shape recognition without a preliminary segmentation process. It is based on a pyramidal model, and exploits the pixels that do or do not cover the shape that makes up the model.

● Use of the ICRPad System

ICRPad provides scholars with the following advantages/functionality:

1. connecting in real time to several existing databases, using the “repository selection” functionality of the “system setting” interface;
2. exploring the images stored in the connected databases to evaluate which items are to be selected to create shape models to be used as search keys;
3. changing at any moment the parameter settings in order to customize the searches and fine-tune the quantity and quality of the results, depending on how much data the user expects to find in the search (deformation thresholds, etc.): the higher the

thresholds combined with the deformation parameters, the more exact the search for shape model occurrences;

4. creating in real time shape models tailored on the user's needs: after displaying one or many images, the user can select, by pointing and zooming directly in the images, the regions he is interested in and create the shape model according to his needs (one graph, many graphs, one word, etc.); an image noise detection tool allows him to check the "dirt" levels that might somehow compromise the reliability of the search;
5. customizing the search by saving selected regions to be used as shape models.

A user-friendly and highly usable interface allows the user to exploit intuitively several tools.

Using the system, first the user selects the "Setting" functionality, that allows him to customize the search parameter settings in order to get results fitting his expectations, or the kind of contents represented in the database(s) of interest, or the graphic item he intends to use as a shape model.

Once the parameter settings have been defined, the user selects from the database(s) he connected for consultation the image(s) from which he wants to create the sample(s).

Then, using the "Create model" functionality, he selects the ROI from the image. It can be any portion of a page image for which he is interested in searching for further occurrences in one or many available database(s) of images: a grapheme, or a glyph, or part of a lemma or a whole lemma, a single line or even a set of lines, images, illuminations or parts thereof. By the ROI, an automatic and zoom inclusive process creates the models and allows him to define the shape contour in real time. In case he believes that those models can be exploited again in future researches, he can save them in a customizable system repository.

Finally, the user moves to the "Find" functionality, and runs his search on the entire set of databases previously selected.

3 A Technical View of the System

The matching algorithm of ICRPad is based on shape contours, even of different size. As a starting point of the process, it extracts a graphic region (an image or a part of it) and creates a shape model to be used for searching similar patterns in one or more target images. After defining the model, in order to obtain satisfactory results, the system must return:

- the position within the document on which the search is performed;
- the angle;
- the scale of the image part found with respect to the specimen given in input;
- a score indicating how much (a percentage) the result is similar to the starting model.

It is possible to handle settings using both a well-rounded set of *primitives* and graphic features as basic parameters for searching portions of the graphic regions in digital images. Some main graphic features managed by the algorithm are: angle, rotation, scale, overlay, contrast, brightness, color, transparency, focus, distortion, occlusion, deformation.

The system embeds a shape-based pattern matching algorithm which recognizes and represents objects by their shape, ignoring their size and the gray-scale values of pixels and their neighborhood in the model. It can be classified as a *spotting* algorithm, since it does not need a preliminary segmentation of the document pages.

There are several ways to determine or describe the shape of an object. ICRPad extracts the shape by selecting all pixels whose contrast with neighboring pixels exceeds a threshold set by the user. Typically, such pixels belong to the object contour. So, given a shape model, the main task of the matching process is to try and find into the target image its occurrences (all of them, or a maximum number thereof, if initially specified).

In particular, the algorithm allows to specify which pixels belong to the model, to speed up the search by using subsampling, to specify a range of orientations, to specify a range of scale and so on.

Finally, the system allows to search many models at the same time within each image, and to parallelize all of the processing for one or more models, in order to optimize the search times by a computational point of view.

Depending on the parameter settings, the process can provide the following features for each model found:

- position, skew angle and score of the found model(s);
- position, skew angle, a uniform resolution factor and the score of the model(s) found;
- position, skew angle, different (horizontal/vertical) resolution factors and score of the fitted model.

If the search concerns several models, information about which model each found instance refers to is also provided.

3.1 Model Creation

The shape model characterizes and defines an internal representation of the portion of image that should be used as a search criterion. This image should be shown in its ideal form, i.e., the sharpest possible, without occlusions and possibly aligned with the horizontal axis.

The source image format to define the model can be any of the common electronic formats, such as TIFF, BMP, GIF, JPEG, PPM, PGM, PNG, PBM, and so on. The image for creating the model may be of any shape (elliptical, circular, polygonal or even outlined freehand) and have arbitrary angle.

Figure 1 shows how models are created. The region surrounding the model is the ROI. The search process optimization starts from the definition of a good model. After defining the portion of image to be used as a model, some of its parameters for the search process may be changed; also, a model can be stored on disk, in order to retrieve and modify it later for future use.

To obtain a suitable model, the contrast must be chosen so that the pixels that are *significant* to the object are included in the model. By *significant pixels* we mean those pixels that characterize the object and allow to clearly discriminate the shape to be searched from other objects and from the background. The model must have minimum noise or a low number of non-interesting regions (i.e., regions not belonging to the object

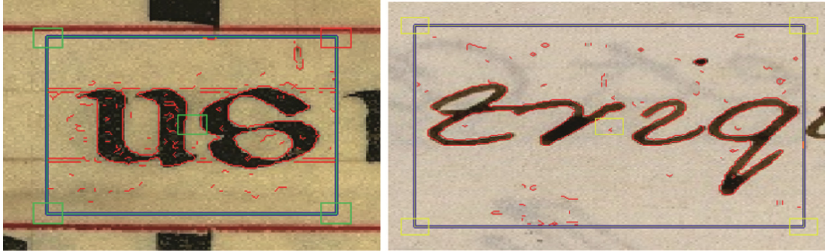


Fig. 1. Creation of models

to be searched). In some cases there is no single contrast threshold that allows to both discriminate the noise and, at the same time, remove non-interesting parts from the graphic region.

3.2 Search Parameters

The most important parameters for searching a model are:

- *contrast*: through the definition of a threshold (low-high) it allows to discriminate the pixels belonging to dirty or irrelevant portions of the image;
- *number of pyramid levels* that make up the model, i.e., the image set consisting of a number of graphic models having different resolution, as shown in Fig. 2: if the original image has a resolution of 600×400 dpi, the pyramid will consist of a first-level image at 600×400 dpi, a second-level image at 300×200 dpi, a third-level image at 150×100 dpi, and so on; this is a crucial factor for performance and accuracy of the results; as general rule, it's a good result if a region of interest has a width of $2^{\text{LevelNumber}} - 1$ pixels (e.g. 8 pixels width allows one to use four pyramid levels); then, after an appropriate region has been set, the reduced image may be used as a model for creating the reference shape;

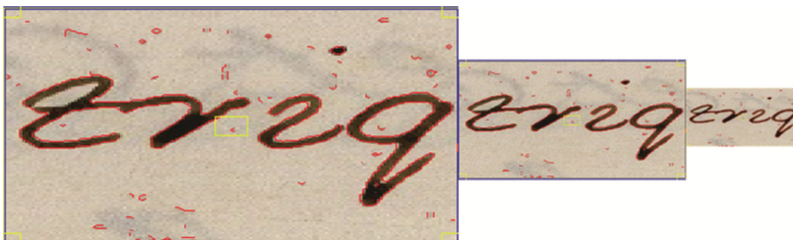


Fig. 2. Number of pyramid levels

- *rotation angle and extension of the model*: e.g., setting the angle to 5° and the extension to 10° , the search could be performed using images with a rotation tolerance of about $\pm 5^\circ$;

- *scale parameter for factors x, y* : this allows to define the pair min/max for each axle in order to run a stretch of the model;
- *timeout parameter*: this allows to speed the search process up to 10%; this is useful when one wants processing to stay within a certain amount of time for each image-destination.

3.3 Configuration Parameters

The various configuration parameters were tested on a dataset of different images, and were calibrated on some test cases by varying gradually the percentage of recognition, in order to improve the steps of the algorithm.

The parameters used during this trial and in the subsequent final definition of the steps of the algorithm are:

- basic:
 - minimum score: a similarity measure between the model to be searched for and a candidate occurrence; the larger the score value, the faster the search, because candidates are discarded earlier; experiments lets us conclude that, given a positive sample, i.e. documents certainly containing the shape model to find, the percentage of recognition required to define the optimal score for a uniform image type are about 90%: such condition means that the parameters defining the model and those for the algorithm execution ensure the expected result;
 - maximum number of items found per image: this is the value to retrieve all potential models from each image.
- advanced (these are essential for the search):
 - completeness: determines the trade-off between efficiency and effectiveness of the search results. A low value determines a complete, but quite slow search; the higher the value, the faster the search, but at the expenses of completeness (i.e., an occurrence of the model might not be found even if it is visible in the image);
 - overlap: specifies how two graphic regions may overlap an image; in case of symmetry, the allowed overlap should be reduced to prevent multiple matches on the same object;
 - sub-pixel: determines accuracy by selecting the precision in calculating the position, orientation and scale; default definitions may be provided for these features, such as calculation of position, which can be determined with precision 'pixels' only, while accuracy of orientation and scale is respectively equal to values of angle and of scale size specified during the construction of the model; in this way, the position is estimated with accuracy at pixel level, and the size of the object determines the accuracy of estimated scale and orientation: the larger the size, the more precise the orientation and scale;
 - deformation: sometimes, the objects are either not found at all or found with a low precision degree only, because they are slightly deformed compared to the model. In these cases, it is possible to use a deformation parameter that expresses how many pixels of deviation are tolerated from the edges found in the image to those of the model shape. The value of this parameter should be set as small as possible,

using a high value only for targeted searches. Indeed, the higher this value, the greater the risk of finding wrong model instances; moreover, a high value for this parameter often produces an increase in processing time; both problems mainly stem from the search of small/fine/thin structured objects, because these kind of objects, when undergoing further deformation, lose their characteristic shape, which is important for an effective search.

3.4 Searching Shape Model

The position and rotation of the found instances are returned as *row*, *column* and *angle* values. Moreover, each instance found is marked by a *score*. Additional information is returned, such as *scale*: if the shape model is creating, the resolution ratio between the model and the found image is parameterized.

Downsampling can be enabled to speed up the matching process, i.e., images at lower resolution can be used. The set of images at different resolutions representing the same source image, ordered by decreasing resolution, is called a *pyramid*, and the images in a pyramid are the pyramid *levels*, where the top of the pyramid is the image at the lowest resolution. When defining a model, a set of images having different resolution is created. So, the model is created and searched on multiple pyramid levels (images). The number of levels of the pyramid to be used can be specified: it is a good practice to choose the highest level of the pyramid with models containing at least 10 to 15 pixels and such that the shape model still resembles the shape of the object.

However, the system provides primitives that allow the user to automatically set these parameters by an internal analysis of the region covered by the model.

4 Experimental Results

The full matching process of ICRPad was tested on about 3500 images belonging to 7 medioeval latin manuscripts dated between the XI and XIII Centuries, all of unknown – but seemingly different – authors and scriptoria.

The graph “&” was used as a sample of this first experiment, because it was distinctive of all researched manuscripts.

Before starting the matching process, we created the shape model. Then we set the deformation parameter and the different resize and minimum score parameters by which searching the patterns within the set of images.

By varying the ratio between deformation, minimum score and resize we got different results for each sample image and thus we could evaluate the effectiveness of each performed pattern.

We had best result with deformation 3, resize 40% and minimum score between 60% and 80%. By setting these parameters we had a high level of True Positives in four manuscripts (numbered 1, 2, 4 and 5): about 80%, with few False Positives in manuscript 5 setting minimum score at 60% (weak/low parameter). Surprisingly, we knew that really these were written either by the same hand or by the same scriptorium. Furthermore, we noticed that some False Positives were to be considered as further evidence of

omography of the four manuscripts, because by overlapping them with some similar True Positives (e.g. C and O, S and F) we saw that some graph curves were identical.

The above parameters can be changed and set either before the search by looking at the shape model that has been created, or during search if the result is not as expected.

5 A Data Science Perspective on Scholarly Research on Data Humanities

We use ICRPad to envision a method that allows scholars to search different types of digital libraries, querying them according to an “assumption-free” approach. It means that the system does not deal with a single, specific, pre-defined database, but it can be connected in real time to several databases available on-line. The user will select one of them as relevant to his research objectives. After selecting the databases, the user chooses the image(s) that will provide the models to be searched. Then, he creates the shape models for his search and, by using them, he queries the connected databases. In the classical setting, he starts from a search hypothesis and checks whether the query results may confirm it. In the new setting we envisage, he may also issue a random search, i.e., without any expectation on the results, and draw inspiration to build new research hypotheses from the very search results. Finally, he might query the databases with a research hypothesis in mind, but in addition to the True Positive outcomes, he may also carefully analyze False Positive outcomes: the latter, in fact, may be of great interest, because some of them, albeit different than the model, may nevertheless reveal similarities that suggest the scholar interesting hypotheses to be investigated. This is a data science approach to the consultation of databases, and opens new frontiers to the study of data humanities.

We show this with the following sample scenario. We tested ICRPad working with a scholar involved in paleographic and historical studies on handwritten codices. He aims at exploring new research hypotheses concerning his domain of interest. To this purpose, he chooses to query some existing on-line databases in order to collect useful hints for his research inspired by the results of his queries. Using ICRPad, he connects to a registered on-line database (e.g., the digital library of The British Library), because he wants to explore two manuscripts included in a famous codex (let us call them “*ms A*” and “*ms B*”), that are generally considered in the literature as written by different writers. He aims at checking whether there is some chance that, using the matching algorithm, the alternate graphs, on which the claim that the two writers are different is based, have in fact sufficient similarity to be considered as written by the same writer. In real time, he connects to the repository in which the codex is stored, chooses at random a page image from one of the two manuscripts under consideration (say, *ms A*), and from this image he selects the greek alphabet glyph ψ as Region of Interest (ROI). Then, he proceeds by creating from the ROI the shape model to search using the “create model” tool. He saves the shape model just created in his personal system repository, then he runs the “find” functionality. The system scans the whole codex and returns all occurrences that fulfil the parameters, and thus are considered equal or similar to the shape models he created. He gets the following results:

- ms A, glyph ψ :
 - True Positives (homograph of ψ): 75% (50% in ms A, 25% in ms B)
 - False Positives: 25% (10% in ms A, 15% in ms B); among which:
 - glyphs nearly homograph of ψ : 20% (5% in ms A, 15% in ms B), all referred to glyph ψ , whose strokes perfectly overlap the corresponding traits of ψ ;
 - glyphs approximately homograph: 5% (all in ms B), all glyphs υ , some traits of whose curved lines can be overlapped to the corresponding traits of ψ .

The True Positive results, albeit uncertain, could be somewhat expected by the user when issuing his search, but might also have been insufficient, alone, to back the hypothesis that the same writer wrote both manuscripts. On the other hand, False Positive results, albeit unexpected, may be even more important, because they turn out to be in many traits totally homograph to the model. In such a case, an item that would normally be considered as a noisy result, may be interpreted as a further confirmation that both manuscripts were made by the same writer. In turn, this stimulates additional investigation on both the digital and the physical artifact, and becomes the starting point for a new hypothesis about the authorship of the two manuscripts. E.g., contrary to what was claimed and believed along many years of literature, it suggests that the two manuscripts were actually manufactured by the same person, operating in one scriptorium or, in different moments, in several scriptoria that used exactly the same “canone”. Another possibility is that the “canone” stayed unchanged during one or two centuries in the same scriptorium or in different scriptoria in the same geographic area, and that it was used according to strict, precise rules by different writers.

6 Conclusion

In this paper we outlined the features of ICRPad, a patented graphic matching system for the digital recognition of manuscripts, and proposed a new approach to searching and retrieving information in digital libraries. This approach is based on applying data science, the fourth paradigm of knowledge development in the scientific field, that is at the basis of science informatics, to studies of data humanities.

The training process is based on the outlined matching algorithm, which uses the recognition of contours shape without any segmentation process. It does not use the gray-scale values of the image. It is based on a pyramidal model, and exploits the pixels that do or do not cover the shape that makes up the model: i.e., the set of images is composed by models at different graphic resolution. This factor is crucial for performance and accuracy of the results because, after setting an appropriate region, the reduced image can be used as image-model for the creation of the shape-model.

The latest system version that we have implemented processes about 240.000 images/h with 60%–90% of positive matches, starting the search from four base minimum score parameters of 60% (weak), 70% (low), 80% (medium), 85–90% (high).

References

1. Barbuti, N., Caldarella, T.: An innovative character recognition for ancient book and archival materials: a segmentation and self-learning based approach. In: Agosti, M., Esposito, F., Ferilli, S., Ferro, N. (eds.) IRCDL 2012. CCIS, vol. 354, pp. 261–270. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-35834-0_26
2. Fischer, A., Bunke, H.: Character prototype selection for handwriting recognition in historical documents. In: Proceedings of 19th European Signal Processing Conference, EUSIPCO, pp. 1435–1439 (2011)
3. Indermühle, E., Eichenberger-Liwicki, M., Bunke, H.: Recognition of handwritten historical documents: HMM-adaptation vs. writer specific training. In: Proceedings of 11th International Conference on Frontiers in Handwriting Recognition, Montreal, Quebec, Canada, pp. 186–191 (2008)
4. Bulacu, M., Schomaker, L.: Automatic handwriting identification on medieval documents. In: 14th International Conference on Image Analysis and Processing, ICIAP 2007, pp. 279–284 (2007)
5. Rath, M.T., Manmatha, R.A., Lavrenko, V.: Search engine for historical manuscript images. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 369–376 (2004)
6. Srihari, S., Huang, C., Srinivasan, H.: A search engine for handwritten documents. In: Document Recognition and Retrieval XII, vol. 154, no. 3, pp. 66–75 (2005)
7. Fischer, A., Wüthrich, M., Liwicki, M., Frinken, L., Bunke, H., Viehhauser, G., Stolz, M.: Automatic transcription of handwritten medieval documents. In: Proceedings of 15th International Conference on Virtual Systems and Multimedia, pp. 137–142 (2009)
8. Adamek, T., O'Connor, E.N., Smeaton, A.F.: Word matching using single closed contours for indexing handwritten historical documents. *Int. J. Doc. Anal. Recogn. (IJ DAR)* **9**(2–4), 153–165 (2007)
9. Herzog, R., Neumann, B., Solth, A.: Computer-based stroke extraction in historical manuscripts, manuscript cultures. *Newsletter* **3**, 14–24 (2011)
10. Krtolica, R.V., Malitsky, S.: Multifont optical character recognition using a box connectivity approach (EP0649113A2) (2012). http://worldwide.espacenet.com/publicationDetails/biblio?CC=EP&NR=0649113&KC=&FT=E&locale=en_EP. Accessed 20 May 2012
11. Leydier, Y., Le Bourgeois, F., Emptoz, H.: Textual indexation of ancient documents. In: Proceedings of the 2005 ACM Symposium on Document Engineering, pp. 111–117 (2005)
12. Dalton, J., Davis, T., van Schaik, S.: Beyond anonymity: paleographic analyses of the Dunhuang manuscripts. *J. Int. Assoc. Tibet. Stud.* **3**, 1–23 (2007)
13. Le Bourgeois, F., Emptoz, H.: DEBORA: Digital AccEss to BOoks of the RenaissAncE. *IJDAR* **9**(2–4), 193–221 (2007)
14. Bar-Yosef, I., Mokeichev, A., Kedem, K., Dinstein, I.: Adaptive shape prior for recognition and variational segmentation of degraded historical characters. *Pattern Recogn.* **42**(12), 3348–3354 (2008)
15. Gordo, A., Llorenz, D., Marzal, A., Prat, F., Vilar, J.M.: State: a multimodal assisted text-transcription system for ancient documents. In: Proceedings of 8th IAPR International Workshop on Document Analysis Systems, DAS 2008, pp. 135–142 (2008)
16. Cheriet, M., et al.: Handwriting recognition research: twenty years of achievement... and beyond. *Pattern Recogn.* **42**, 3131–3135 (2006)
17. Le Bourgeois, F., Emptoz, H.: Towards an omnilingual word retrieval system for ancient manuscripts. *Pattern Recogn.* **42**(9), 2089–2105 (2009)

18. Nel, E.-M., Preez, J.A., Herbst, B.M.: A pseudo-skeletonization algorithm for static handwritten scripts. *Int. J. Doc. Anal. Recogn. (IJ DAR)* **12**, 47–62 (2009)
19. Stokes, P.A.: Computer-aided palaeography, present and future. In: Rehbein, M., et al. (eds.) *Codicology and Palaeography in the Digital Age*, Schriften des Instituts für Dokumentologie und Editorik, Band 2. Book on Demand GmbH, Norderstedt (2009)
20. Toselli, A.H., Romero, V., Pastor, M., Vidal, E.: Multimodal interactive transcription of text images. *Pattern Recogn.* **43**(5), 1814–1825 (2010)
21. Fischer, A., Wüthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G., Stolz, M.: Automatic transcription of handwritten medieval documents. In: *Proceedings 15th International Conference on Virtual Systems and Multimedia*, pp. 137–142 (2009)

XDOCS: An Application to Index Historical Documents

Federico Bolelli^(✉), Guido Borghi, and Costantino Grana

Dipartimento di Ingegneria “Enzo Ferrari”,
Università degli Studi di Modena e Reggio Emilia,
Via Vivarelli 10, 41125 Modena, Italy
{federico.bolelli,guido.borghi,costantino.grana}@unimore.it

Abstract. Dematerialization and digitalization of historical documents are key elements for their availability, preservation and diffusion. Unfortunately, the conversion from handwritten to digitalized documents presents several technical challenges.

The XDOCS project is created with the main goal of making available and extending the usability of historical documents for a great variety of audience, like scholars, institutions and libraries. In this paper, the core elements of XDOCS, *i.e.* page dewarping and word spotting technique, are described and two new applications, *i.e.* annotation/indexing and search tool, are presented.

Keywords: Indexing · Page dewarping · Word spotting
Word annotation · Handwriting recognition

1 Introduction

The availability of large collection of handwritten historical manuscripts is often required and craved by libraries, scholars and institutions. Despite this, many issues are related to these particular documents.

First of all, the diffusion of historical documents is strictly limited by their physical condition and, often, their are available in a single copy. Moreover, the document readability can be compromised due to the presence of particular handwriting style or other graphic artifacts belonging to old writing techniques. A solution for these problems can be represented by the dematerialization and digitalization of documents and, in this context, the creation and collection of the so called *Digital Libraries* [1, 5, 13] represents a key elements in the process of diffusion, usability and availability of historical documents. The XDOCS project is designed with the intention of extending the audience and the access of a huge variety of Italian historical documents.

The conversion from handwritten to digitalized documents represents a great challenge from a technical point of view. On one hand, the peculiarity of this kind of data and the huge amount of documents exclude the possibility to exploit manual annotations and operations which are extremely time-consuming and

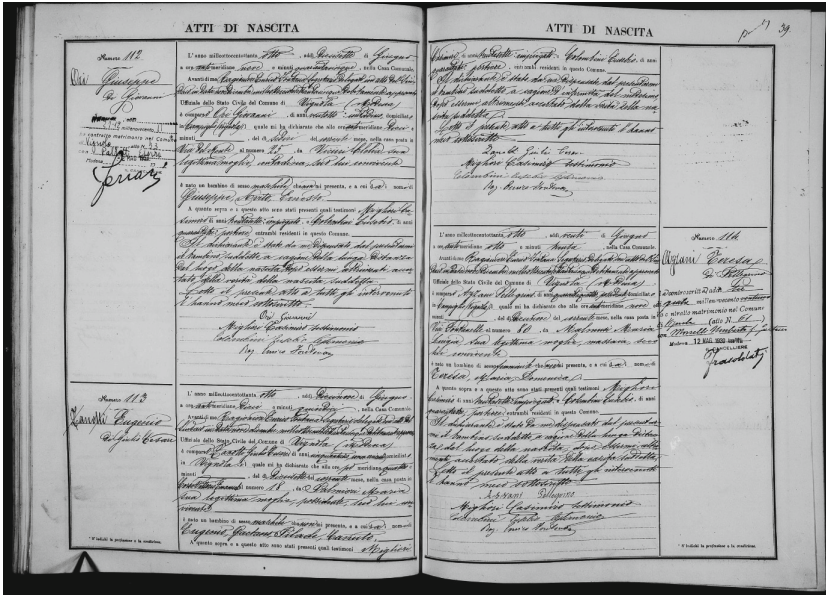


Fig. 1. Example of historical document page dated 1888 and representing three birth acts of the Italian state.

expensive. On the other hand, classical automatic writing recognizers, also called *Optical Character Recognizers (OCRs)*, often fail.

One of the first step after the digital acquisition of the document is the page dewarping, since warping distortion affects, as well as the document readability, the performance of automatic techniques of content mining, indexing and word annotation. The input of this process is represented by a curled page (see for instance Fig. 1), usually captured by a flatbed scanner and the output is a page containing only horizontal straight lines, without any distortion due to perspective or page warping.

Once historical documents are correctly digitalized and dewarped, it is possible to apply word annotation and word spotting techniques to facilitate the study of researchers and the extraction of semantic contents. A word spotting technique is the ability to create word collections grouped into clusters containing all instances of the same word. Exploiting this technique, it is then possible to index in a semi-automatic way the content of documents.

The paper is organized as follows. Section 2 presents an overall description of related literature works, divided into two main group: page dewarping and word spotting tasks. In Sect. 3, a general overview of the XDOCS project is given with particular attention to dewarping and word spotting techniques. Section 4 describes the annotation/indexing and search tools. All datasets exploited in the project are described in Sect. 5. Finally, in Sect. 6 conclusions are drawn.

2 Related Work

Page dewarping. Over the last two decades, many document dewarping techniques have been proposed. The main issue of those proposals is that they are specifically designed for typewritten text, so, they produce low quality results when applied to handwritten texts or hybrid documents (documents that contain a mix of typewritten and handwritten text). Generally, we can divide these approaches in two categories according to the surface model adopted: restoration approaches based on 3D document shape reconstruction [4, 7] and restoration approaches based on 2D document image processing [8, 20]. The 3D reconstruction models are more accurate but they require images captured with special setup to properly work and this is not the case of common historical digital documents. The 2D approaches, instead, make use of the information contained in the document image in order to restore the page, so they are much more suitable for historical documents. An interesting technique belonging to the second group has been proposed by Stamatopoulos *et al.* in [16]: a two-step approach for efficient dewarping. At the first step, a coarse dewarping is completed with the help of a transformation model, in which a curved surface is projected in a 2D rectangular area. At the second step, fine dewarping is conducted thanks to the word detection, since all words poses are normalized based on the lower and upper word baselines. In [2] a novel approach based on [16] for performing dewarping on Italian historical document images, containing both typewritten and handwritten texts, is presented. This represents the baseline of the XDOCS project so it is described in details in Sect. 3.1.

Word Spotting. In [11, 12], the original idea of word spotting for handwritten manuscripts was proposed. In these works the matching techniques and pruning methods are described: given a word image, similar words are clustered and unlikely matches are quickly discarded. Generally, word spotting methods can be divided in two main categories: *line-segmentation* and *word-segmentation* based approaches.

Word-segmentation approaches are based on the hypothesis that each word in the document images is separately clipped. Tomai *et al.* [19] proposed a word-by-word mapping between a scanned document and a manual transcript: in this way, it is possible to perform word location in document pages. This method relies on a *Optical Character Recognizer* (OCR) used as a recognizer for multiple word segmentation hypothesis generated for each line of the document. Results shown that a OCR is not a feasible solution and useful for handwritten historical manuscript recognition. In [15] a local descriptor inspired by a famous key-point descriptor, SIFT [10], is proposed. Here, two different word spotting systems, based on the well-known *Hidden Markov Models* and *Dinamic Time Warping* (DTW), are exploited to achieve significant improvements. In [14] a range of features suitable for DTW has been analyzed. In that paper, different text features, which are extracted from pre-processed rectangular word images and that do not contain ascenders from other words, are used to achieve speed and precision. Exploited features are the gray scale variance,

the projection profile, background to ink transitions, the partial projection profile, the upper and lower word profile, and feature sets containing vertical and horizontal partial derivatives. All of them were extracted after normalization of inter-word variations such as *skew* and *slant* angles.

Line-segmentation based methods rely on the hypothesis that each line in the document is separated and word segmentation techniques are not strictly required. Terasawa *et al.* [17, 18] presented a word spotting method based on sliding window, line segmentation, continuous dynamic programming and a gradient-distribution-based feature with overlapping normalization and redundant expressions.

In [9] a line-oriented process is applied to avoid the problem of segmenting cursive script into individual words. This approach exploits dynamic programming algorithms and pattern matching techniques. The proposed system is tested on old Spanish manuscripts, showing a high recognition rate. Unfortunately, it is much expensive since words have to be searched for every possible position. Besides, DTW is separately applied on feature vectors and results are merged, producing different alignment for the same word-line pair.

3 The XDOCS Project

Due to the great amount of variability in handwriting styles and the high noise levels in historical documents, handwritten historical documents are generally transcribed by hand. The main goal of the XDOCS project is to develop an innovative data capturing technique able to extract document indexes in quasi-automatic mode from their handwritten contents in order to extend the usability of the historical documents. From a general point of view, the XDOCS application could be split into three main blocks. The first one is the *page dewarping* step during which input digital documents are dewarped and normalized. The second one, the *word spotting* phase, aims at building clusters of words with the same index. Finally, the third block of the project concerns words annotation and smart search of indexes inside the historical documents.

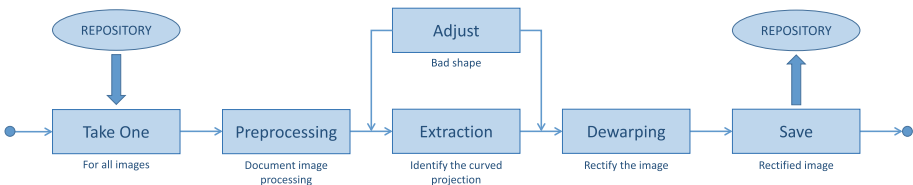


Fig. 2. Pipeline of *Page Dewarping* phase of the XDOCS project.

3.1 Page Dewarping

This step aims to transform the original curled document pages into ones constituted only of horizontal straight text lines, without any distortion caused by lenses and perspective. *Page dewarping*, depicted in Fig. 2, is essentially composed by three steps:

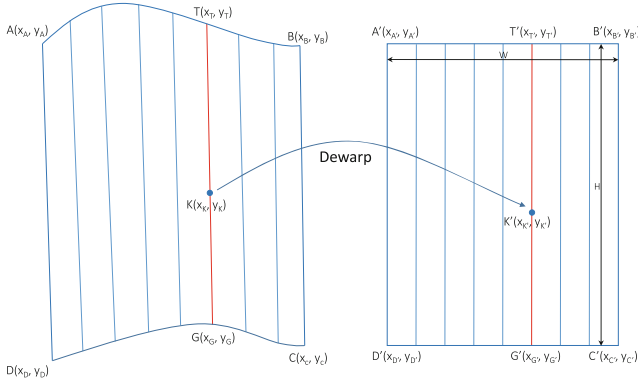


Fig. 3. Dewarping transformation model: projection of the curved surface on the left side, 2D rectangular destination area on the right side.

- *Image pre-processing* step consists in filtering out document and page noises, mainly caused by the intrinsic nature of the original images which belongs to old documents and the digitization process. This filtering relies on connected components statistics as described in [2].
- *Projection extraction* module aims to identify the curved 2D projection surface surrounding the document page. According to the warping model, the projection is assumed to be described by two almost vertical straight lines ($y = ax + b$) and by two third degree polynomial curves ($y = ax^3 + bx^2 + cx + d$). The vertical lines are automatically identified by the use of the *Hough* transform [6], while the coefficients of polynomial lines are fitted with the Least Square Estimation algorithm. Boundary extraction significantly influences the quality of the dewarping process, and then the entire pipeline of the XDOCS application: if it fails the *Adjust* step leaves the user the possibility to correct curves via a GUI.
- *Dewarping* is the core of the image rectification process. During this phase, the projection of the curved surface is mapped into a rectangular normalized 2D area. The transformation is described by Eq. 1 where, referring to Fig. 3, $|AD|$ and $|BC|$ are euclidean distances and $|\widehat{AB}|$ and $|\widehat{CD}|$ are the length of polynomial curves on the projection surface. Moreover, the two points T and G belong respectively to the curves $|\widehat{AB}|$ and $|\widehat{DC}|$. The idea is to preserve proportions between dimensions of projected curves and 2D destination area.

$$K'(x'_A + W * \frac{|\widehat{AT}|}{|\widehat{AB}|}, y'_A + H * \frac{|TK|}{|TG|}) \quad (1)$$

3.2 Word Spotting

At the end of *Page Dewarping*, input images are correctly rectified: they do not suffer of any distortion effects, they are normalized to fixed dimensions and are then ready for the *Word Spotting* [3].

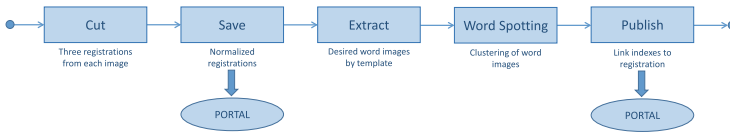


Fig. 4. Pipeline of *Word Spotting* phase of the XDOCS project.

The aim of this step is to group all words of interest into clusters in order to greatly reduce the amount of annotation work that has to be performed in the last phase of the XDOCS project. *Word Spotting* articulates as follows:

- *Cut* three registrations from each image in order to logical separate document contents. This step will produce a series of images like the one reported in Fig. 5. Each resulting image will be stored in a database for user consultation.
- Words representing intended indexes are then *Extracted* exploiting a simple template approach: given that all acts have the same structure (for a give historical book) and were normalized in the previous step, the extraction template can be defined once for all documents.
- *Word Spotting* is the core step of the current process. Firstly, word images are preprocessed and normalized as described in [3]. Then, HOG feature vectors are extracted from each word image exploiting a sliding window approach. Finally, words are matched and grouped together using the similarity distance obtained with DTW technique.

3.3 Indexing and Search Tools

The third part of the XDOCS project is constituted by the annotation/indexing tool and by the advanced search system. Since they are the main novelty of this work they are described in detail in Sect. 4.

4 XDOCS Application

4.1 Annotation/Indexing Tool

The annotation system is based on the word spotting approach described in the Sect. 3.2. When a new registry is loaded onto the system it is processed following the pipeline of Fig. 4 and all extracted words associated to the same index are compared together. After that, it is possible to build a ranking system and store into the database all matches between a word and those most similar to it. The association list can be consulted by the users, as shown in Fig. 6. The tool provides, for each word/index the following information:

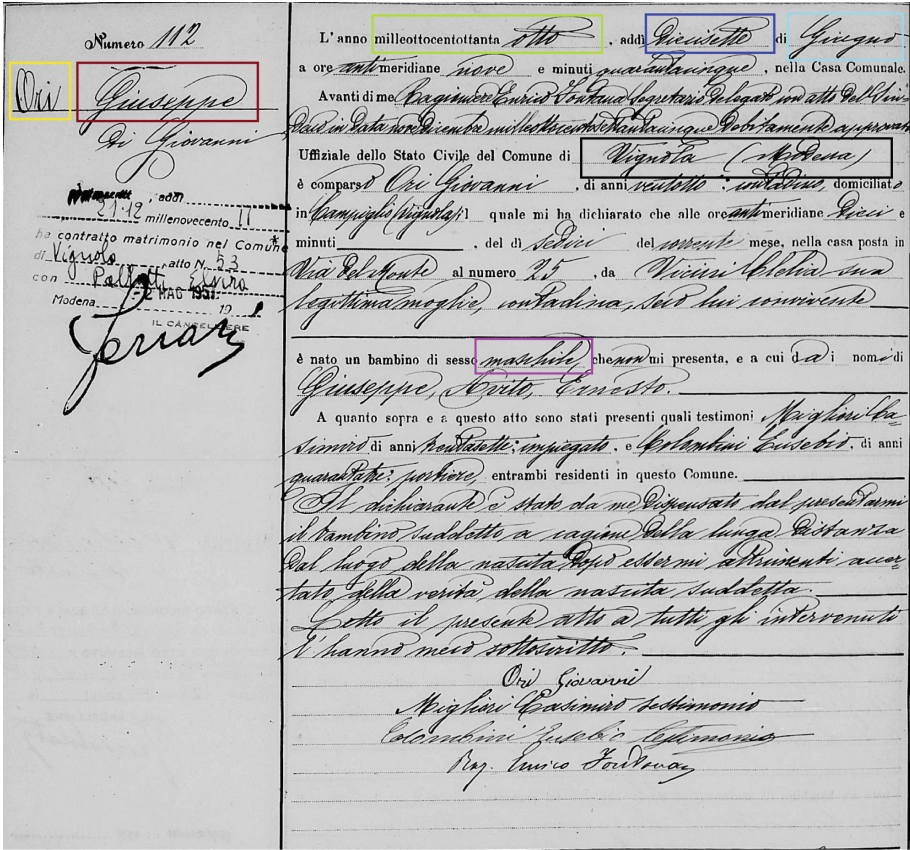


Fig. 5. Example of single birth act registration obtained after the *Cut* step of *Word Spotting*. The colored bounding boxes contain searched indexes and are automatically extracted by a template approach. (Color figure online)

- General information about the document from which the word belongs to;
- The word image of the index and the associated plain text value, when available;
- A list of most similar words: for each of them, the distance calculated by the word spotting algorithm, the associated image, and a check box to specify if the indexes are actually equivalent are reported.

Thanks to the annotation interface it is possible to set or update information automatically extracted by the *Word Spotting* algorithm and propagate them to the associated words. This procedure allows to significantly reduce time and costs of the annotation process improving the quality and the performance of the search system described in the following section.

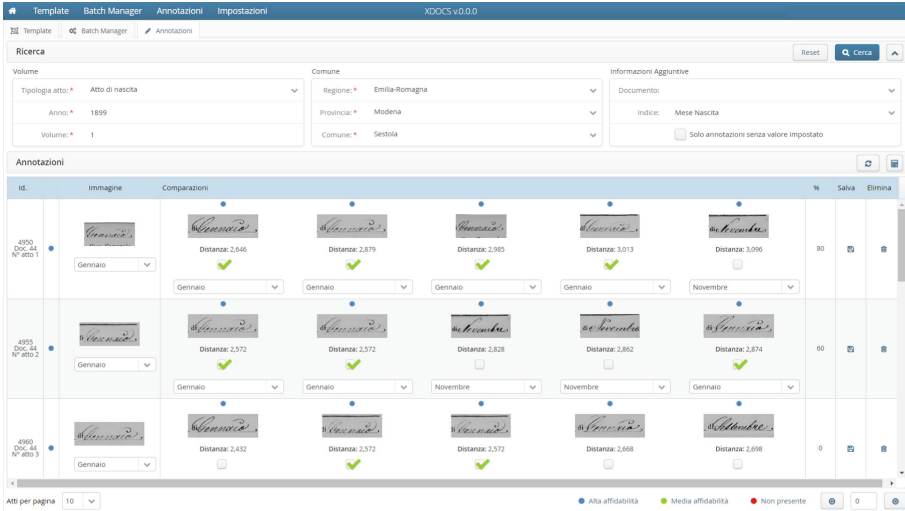


Fig. 6. Annotation interface of the XDOCS project

4.2 Search Tool

In order to achieve the goal of simplifying accesses to historical documents, a browser interface based on *PostgreSQL* database has been developed. The software provides an advanced search tool which allows the user to search single act page specifying, for example, the type (1), the year (2) and the municipality of the searched act (3) (see for instance Fig. 7a). Additional search fields (4) are available after selecting the act type. Figure 7a reports an example of advanced search fields specific for the birth act: name, surname, sex, day/month/year of birth, father name, mother name and so on. It is important to highlight that all search parameters, but the act type, the year and the municipality, are optional, and the *full text search* of *PostgreSQL* backend will combine them to provide the best search results.

The acts identified by the search process will be displayed in list, ordered by id (Fig. 7b). As depicted in see Fig. 7c, it is possible to select and display one specific act and all the associated words which are classified in three categories distinguished by colors:

- Red: words which require annotation;
- Green: words annotated by user without administrator privileges;
- Blue: words specified by an administrator user.

Additionally, through this interface, it is possible to set and change the value of indexes: any change can be automatically propagated depending on user permissions.

The screenshot shows the search interface of XDOCS v0.0.0. At the top, there is a search bar with 'Ricerca ABI' and a search button. Below it, the 'Ricerca' section contains several filters: 'Volume' (with 'Tipologia atto:*' set to 'Atto di nascita'), 'Da anno:*' (set to '1899'), 'A anno:', 'Comune', 'Regione:*' (set to 'Emilia-Romagna'), 'Provincia:', and 'Comune:'. A red box highlights the 'Ricerca Avanzata per Atti di nascita' section, which includes fields for 'Nome', 'Cognome', 'Sesso', 'Giorno nascita', 'Mese nascita', 'Anno nascita', 'Nome padre', 'Nome madre', 'Cognome madre', 'Nome nonno paterno', and 'Nome nonno materno'.

(a) Advanced search page.

The screenshot shows the search results page of XDOCS v0.0.0. The search filters are the same as in (a). Below the filters is a table with the following columns: 'id', 'Volume', 'Regione', 'Provincia', 'Comune', 'Comune Originario', 'Nome', 'Cognome', 'Sesso', 'Giorno', 'Mese', 'Anno', and 'Apri'. The table contains 10 rows of results, all from the 'Emilia-Romagna' region, 'Modena' province, and 'Settola' commune. The names listed are: Vittoria Cecilia Ferrari, Emmanigildo Zuccarini, Giulia Zuccarini, Rosina Gherardini, Isabella Zecchi, Nicola Ricci, Luigi Gherardini, Ettore Buldrini, Giustino Giuseppe Marchetti, and Cosimo Gherardini. At the bottom, there are filters for 'Alta affidabilità', 'Media affidabilità', and 'Non valutabile', and a 'Atti per pagina' dropdown set to 10.

(b) Search results page.

The screenshot shows the 'Indici' (Indexes) page for a specific birth act. On the left, there is a list of fields with their corresponding values: 'N° atto:' (empty), 'Nome:' (Vittoria Cecilia), 'Cognome:' (Ferrari), 'Sesso:' (F), 'Giorno nascita:' (17), 'Mese nascita:' (Gennaio), 'Anno nascita:' (1899), 'Nome padre:' (empty), 'Età padre:' (empty), 'Nome madre:' (empty), 'Cognome madre:' (empty), and 'Messa in atto:' (empty). On the right, there is a preview of the original document image, which is a handwritten birth act from the 'Comune di Settola'. The text in the image is partially legible and matches the search results shown in (b). At the bottom, there are filters for 'Alta affidabilità', 'Media affidabilità', and 'Non presente', and a 'Dimensione' dropdown set to 50%.

(c) Indexes associated to a birth act.

Fig. 7. XDOCS search tool. (Color figure online)



Fig. 8. Inter dataset variations, *i.e.* the same month name written by different writers.

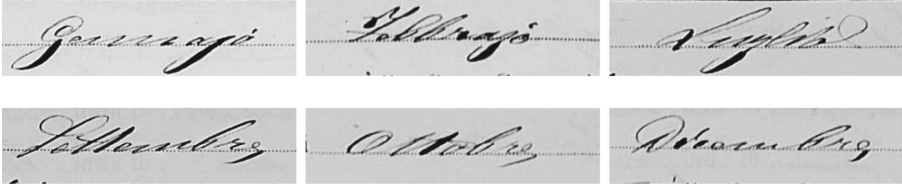


Fig. 9. Intra dataset variations, *i.e.* different months written by the same writer.

5 XDOCS Dataset

As mentioned above, XDOCS is designed with the intention of extending to a much wider audience the possibility to access a variety of historical documents. To that purpose, a great amount of Italian historical birth certificates documents of the XIX century has been collected.

Moreover, to test and evaluate the proposed tools, *Word Spotting* and *Annotation*, a huge collection of single word images has been collected and annotated. All these datasets are publicly released¹.

5.1 Personal Information Data

This is a newly collected dataset consisting of annotated words images of names, surnames, birthdays, municipalities and sex (see Fig. 10 for instance). All images are taken from Italian civil registries of the XIX century. Writing styles variety is guaranteed due to the presence of different writers.

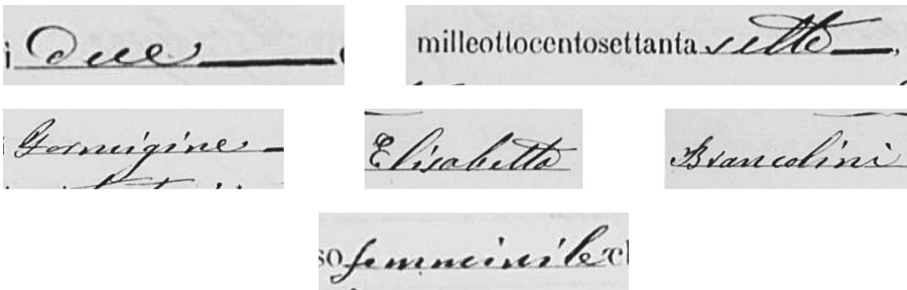


Fig. 10. Examples of intended indexes extract using a template approach from normalized registries. In turn they are: day and year of birth, municipality, name, surname and sex.

¹ <http://imagelab.ing.unimore.it/XDOCS>.

5.2 Months Data

This sub-dataset is firstly introduced in [3] and consists of a collection of handwritten month names extracted from Italian civil registries of the XIX century.

Specifically, the dataset counts 1200 words of all 12 months: data are affected by both *intra* and *inter* variations, due to the presence of three different official state writers, as depicted in Figs. 8 and 9.

6 Conclusions

In this paper we describe the XDOCS project, that includes the page dewarping and the word spotting techniques. Moreover, two frameworks are introduced and described. The first one, the *Annotation* tool, is created to facilitate the annotation of words belonging to historical handwritten documents and the second one, the *Search* tool, is designed to allow searching of these words.

The XDOCS project has the main goal of encouraging the diffusion of handwritten historical documents, generally characterized by difficulties in readability, comprehension and physical availability.

Acknowledgement. The XDOCS project is currently underway at SATA s.r.l. in collaboration with the University of Modena and Reggio-Emilia, and co-funded by the Emilia-Romagna regional administration.

References

1. Balducci, F., Borghi, G.: An annotation tool for a digital library system of epidermal data. In: Grana, C., Baraldi, L. (eds.) IRCDL 2017. CCIS, vol. 733, pp. 173–186. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68130-6_14
2. Bolelli, F.: Indexing of historical document images: ad hoc dewarping technique for handwritten text. In: Grana, C., Baraldi, L. (eds.) IRCDL 2017. CCIS, vol. 733, pp. 45–55. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68130-6_4
3. Bolelli, F., Borghi, G., Grana, C.: Historical handwritten text images word spotting through sliding window HOG features. In: Battiato, S., Gallo, G., Schettini, R., Stanco, F. (eds.) ICIAP 2017. LNCS, vol. 10484, pp. 729–738. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68560-1_65
4. Cao, H., Ding, X., Liu, C.: Rectifying the bound document image captured by the camera: a model based approach. In: Proceedings of the Seventh International Conference on Document Analysis and Recognition, pp. 71–74. IEEE (2003)
5. Corbelli, A., Baraldi, L., Balducci, F., Grana, C., Cucchiara, R.: Layout analysis and content classification in digitized books. In: Agosti, M., Bertini, M., Ferilli, S., Marinai, S., Orio, N. (eds.) IRCDL 2016. CCIS, vol. 701, pp. 153–165. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56300-8_14
6. Duda, R.O., Hart, P.E.: Use of the hough transformation to detect lines and curves in pictures. Commun. ACM **15**(1), 11–15 (1972)
7. Fu, B., Wu, M., Li, R., Li, W., Xu, Z., Yang, C.: A model-based book dewarping method using text line detection. In: Proceedings of the 2nd International Workshop on Camera Based Document Analysis and Recognition, Curitiba, Barazil, pp. 63–70 (2007)

8. Gatos, B., Pratikakis, I., Ntirogiannis, K.: Segmentation based recovery of arbitrarily warped document images. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 2, pp. 989–993. IEEE (2007)
9. Kolcz, A., Alspector, J., Augusteijn, M., Carlson, R., Popescu, G.V.: A line-oriented approach to word spotting in handwritten documents. *Pattern Anal. Appl.* **3**(2), 153–168 (2000)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)
11. Manmatha, R., Croft, W.: Word spotting: Indexing handwritten archives. In: Intelligent Multimedia Information Retrieval Collection, pp. 43–64 (1997)
12. Manmatha, R., Han, C., Riseman, E.M., Croft, W.B.: Indexing handwriting using word matching. In: Proceedings of the first ACM International Conference on Digital Libraries, pp. 151–159. ACM (1996)
13. Pini, S., Cornia, M., Baraldi, L., Cucchiara, R.: Towards video captioning with naming: a novel dataset and a multi-modal approach. In: Battiato, S., Gallo, G., Schettini, R., Stanco, F. (eds.) ICIAP 2017. LNCS, vol. 10485, pp. 384–395. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68548-9_36
14. Rath, T.M., Manmatha, R.: Features for word spotting in historical manuscripts. In: Proceedings of the Seventh International Conference on Document Analysis and Recognition, pp. 218–222. IEEE (2003)
15. Rodriguez, J.A., Perronnin, F.: Local gradient histogram features for word spotting in unconstrained handwritten documents. In: Proceedings of the 1st ICFHR, pp. 7–12 (2008)
16. Stamatopoulos, N., Gatos, B., Pratikakis, I., Perantonis, S.J.: A two-step dewarping of camera document images. In: The Eighth IAPR International Workshop on Document Analysis Systems, DAS 2008, pp. 209–216. IEEE (2008)
17. Terasawa, K., Nagasaki, T., Kawashima, T.: Eigenspace method for text retrieval in historical document images. In: Proceedings of the Eighth International Conference on Document Analysis and Recognition, pp. 437–441. IEEE (2005)
18. Terasawa, K., Tanaka, Y.: Slit style hog feature for document image word spotting. In: 10th International Conference on Document Analysis and Recognition, ICDAR 2009, pp. 116–120. IEEE (2009)
19. Tomai, C.I., Zhang, B., Govindaraju, V.: Transcript mapping for historic handwritten document images. In: Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition, pp. 413–418. IEEE (2002)
20. Ulges, A., Lampert, C.H., Breuel, T.M.: Document image dewarping using robust estimation of curled text lines. In: Eighth International Conference on Document Analysis and Recognition (ICDAR 2005), pp. 1001–1005. IEEE (2005)

Object Recognition and Tracking for Smart Audio Guides

Lorenzo Seidenari^(✉), Claudio Baccchi, Tiberio Uricchio, Andrea Ferracani,
Marco Bertini, and Alberto Del Bimbo

University of Florence, Florence, Italy
{lorenzo.seidenari,claudio.baecchi,tiberio.uricchio,
andrea.ferracani,marco.bertini,alberto.delbimbo}@unifi.it

Abstract. In this paper we address the problem of creating a smart audio guide that adapts to the actions and interests of tourists. Our guide performs automatic recognition of artworks and allows the users instant or deferred fruition of multimedia content. We use a compact CNN as computer vision system to back the whole application to performs object classification, localization and recognition. Tracking is used to improve the recognition accuracy over sequences of detections. We also provide an automatic pipeline for dataset creation based on the same tracking algorithm. The system, deployed on an NVIDIA Jetson TK1 and an NVIDIA Shield Tablet, has been tested in a real world environment.

Keywords: Object recognition · Cultural heritage

1 Introduction

According to recent statistics from the US National Travel and Tourism Office, a new record of tourism-related activities¹ has been set recently. Museum visits are rising steadily thanks to the availability of new digital and mobile technologies. Modern visitors do not follow fixed paths, but they expect personalization and interaction. As a result, new companion tools are needed, providing content sized to the needs and interests of the visitors [1].

In order to automatically gather the behavior of users, these tools have been using cameras to observe where the users go and what they observe. Several approaches resorted to computer vision systems to offer recommendation based on passive external behaviour observation [2] or, more recently, to develop a wearable smart audio guide [3] to automatically play content or interact with artworks [4]. These modern approaches work by constantly matching the user point of view with a visual database of the known artworks, deciding, depending on user behavior, whether to start or not the audio description generated by means of text to speech technology [3] or to show additional content on gestures [4]. Although designed to work in different settings, they all require a computer

¹ <http://tinet.ita.doc.gov/tinews/archive/tinews2017/20170413.asp>.

vision expert to train and test a computer vision models for artworks, person or statues. Moreover, every time an artwork is added or removed, the database has to be updated and a new model has to be trained. We argue that a more efficient solution would be to let the museum curator add new artworks by himself, without requiring to retrain the model from scratch.

In this paper we propose a wearable audio guide system that, by observing in real time what the user is looking at and by following him in the visit, provides personalized content when needed. The device observes the wearer behavior through a computer vision system and decides when to start and stop the reproduction of the audio content. In contrast to previous work [3], artworks are recognized from an on-board database that can be easily made by museum curators using a novel, easy to use procedure. To this end, we show how to avoid re-training the object detector by learning a generic artwork detector based on convolutional neural network (CNN). We develop a novel artwork tracking technique that is used to populate the database using the same CNN object detector we trained for recognition.

We implemented an Android application that a museum curator can use to build a dataset of artworks adaptively. After a recording phase, it performs all required computation on board and outputs ready to use models for in smart audio guides like [3].

2 Related Work

Our work is mainly related to the personalization of the cultural experience and content recommendation on mobile devices. Many works propose to use mobile systems to enjoy an augmented personalized experience on cultural heritage. One of the first concept was that of Abowd *et al.* [5], that marked the difference between indoor and outdoor systems. We thus follow that division and differentiate between local systems to be used in cultural heritage sites, where there is control over the artworks, and outdoor systems that can be used while traveling in a city.

Local systems are mostly developed for museums. In [6] the Cultural Heritage Information Personalization (CHIP) system was proposed, where a personalized tour could be created through a web interface. The tour can be downloaded to a mobile device using RFID present in the museum, and keeps track of the visited artwork on the server side user profile for successive tours. Analyzing and predicting the behavioral patterns of museum visitors, through the use of interactive digital guides was proposed in [7, 8]. They follow the four identified patterns, emerged through ethnographic observations in [9]. Augmented reality on a mobile device was explored in [10] to offer a personalized interactive storytelling experience. Based on the age of the visitor, the system provides a gamified experience to children. In [2] a non-intrusive computer-vision system has been employed to perform re-identification and tracking of users in a museum. By observing the interest of the visitors, it can build a user profile that is then used to create a personalized exploration of multimedia content on an interactive table.

Differently from all of these works, we developed a wearable agent that observes the same scene as the user and provides a novel contextually aware interaction based on audio only, that is unintrusive. Moreover, all the computation required is performed onboard.

3 Efficient Object Detection and Recognition

The smart audio guide we developed is based on an efficient computer vision pipeline that simultaneously performs artwork localization and recognition. The guide requires two main computer vision tasks to be solved: (i) detection of relevant object categories: e.g. persons and artworks; and (ii) for every detected artwork, reliable recognition of the specific artwork framed. Moreover, since we are dealing with a sequence of frames, in order to improve artwork recognition we take advantage from temporal coherence to make the output more stable. Our method is based on [3], which we briefly cover in the following. We use YOLO [11] network that has the main advantage of processing each frame just once to locate all the objects of interest yielding accurate results even for moderate size networks. The architecture is derived from *Tiny Net*, a small CNN pre-trained on ImageNet, which allows the application to run at 10 FPS and fitting on the memory of a Shield Tablet. The system was fine-tuned to recognize artworks and people using our dataset. Recognizing people is relevant for two reasons: first we can exploit the presence of people in the field of view to create a better understanding of context; secondly, without learning a person model it is hard to avoid false positives on people, since artwork training data contains statues which may picture human figures. Learning jointly a person and an artwork model, the network features can be trained to discriminate between this two classes.

3.1 Artwork Recognition

The rich features computed by the convolutional layers are exploited and re-used to compute an object descriptor for artwork recognition. To obtain a low dimensional fixed size descriptor of a region, we apply a global max-pooling over two convolutional feature activation maps and concatenate the result, as shown in Fig. 1. The region is remapped from the frame to the convolutional activation map with a simple similarity transformation. As also detailed in [3], through experimental evaluation, we selected the features from layers 3 and 4, yielding a feature of size 768.

Considering a pre-acquired dataset of artwork patches $p_i \in \mathcal{D}$ and their artwork labels y , for each detected artwork d we predict a specific artwork label $y_{\hat{p}}$ finding the nearest neighbor patch

$$\hat{p} = \arg \max_i \langle p_i, d \rangle \quad (1)$$

The recognition system observes each frame independently and predicts artwork labels according to Eq. 1, this approach, in case of motion blur or quick lighting changes may produce incorrect recognition results.

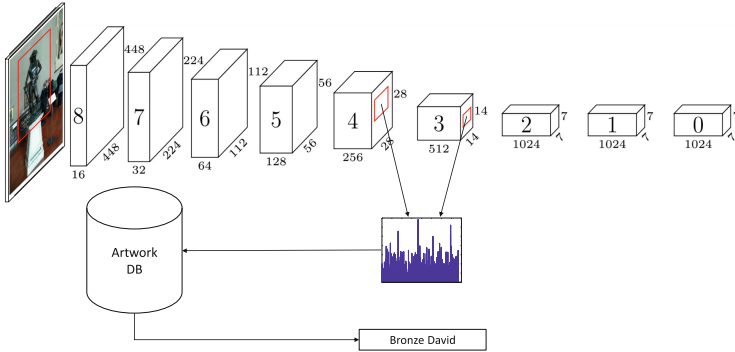


Fig. 1. Our pipeline for recognition, showing network architecture, feature pooling.

4 Automatic Dataset Creation

Extending the dataset with our architecture is extremely straightforward. We rely on a simple multi-target tracking algorithm. With respect to [3] we added a functionality to manage two new use cases: (i) *adding a new artwork*, which is needed in the deployment phase of our system to populate the Artwork DB and whenever a new piece is added to the exhibition; and (ii) *adding examples of an existing artwork*, which arise at any time the position of artworks or any other environmental condition has caused a decrease in performance of the recognition. Moreover, this second use case allows the exhibition curator to acquire artwork samples at multiple times making the acquisition process easier and less fatiguing.

We perform tracking by data association, first we detect all artworks using our CNN, then we greedily associate bounding boxes to the one detected on the previous frame, allowing association only if intersection over union is above 60%. All unassociated boxes are stored and an association is attempted with all boxes at the following frame. All boxes from the previous frame which could not be associated are removed. This method may fail in case the detector skips a frame, nonetheless we found out that this is a very infrequent case and we allow the user to re-initialize the tracker in case the tracked object is lost.

We only retrieve features for an artwork at a time. When the acquisition view is started, the user is prompted to select one of the detected artworks, enclosed in a dashed bounding boxes as show in Fig. 2, from the rolling video. Once an artwork is selected with a tap, its bounding box is drawn with a solid line and a tracking id is shown (for debugging purposes). For every associated detection our CV system stores in the App database the feature extracted using the method described in Sect. 3 together with the relative frame snapshots.

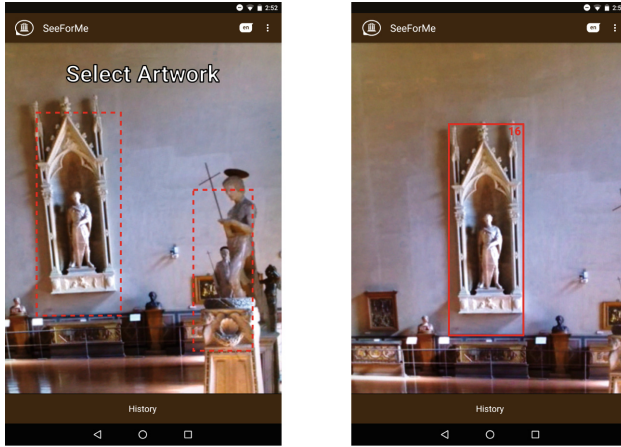


Fig. 2. Tracking process example. User initialization is requested showing all available objects with dashed contours (left). Once tracking has started the tracked object bounding box is shown with solid line and its tracking id (right).

5 Experiments

To show the benefits of our approach for dataset extension in our system, we conduct a simple experiment. We progressively increase the dataset of our artworks reaching a maximum of roughly 2k samples for eight artworks. It can be seen in Fig. 3 that recognition accuracy increases with the amount of samples. It has to be noted that just with the 10% of our acquisition we can reach more than 90% accuracy. Nonetheless increasing the samples reaches almost 100% accuracy. As also shown in [3] the 1-NN approach is best.

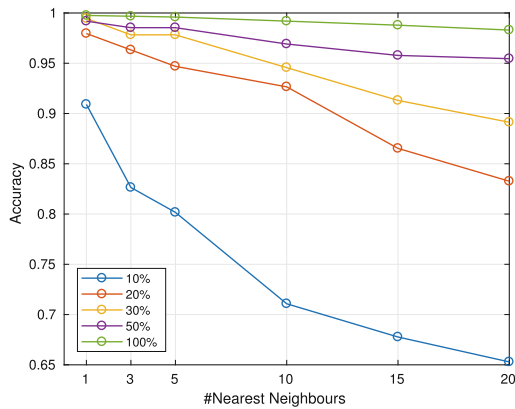


Fig. 3. Accuracy of recognition increasing the dataset size.

6 Conclusions

We have presented a mobile application able to deliver real-time audio information. The main issue of computer vision systems is the training and deployment. We avoid re-training the object detector by learning a generic artwork detector. We show how to populate the database using the same CNN object detector we trained for recognition. Experiments show the benefit of increasing the samples in our pipeline. The system have been deployed and tested on a NVIDIA Shield with TK1.

References

1. Bowen, J.P., Filippini-Fantoni, S.: Personalization and the web from a museum perspective. In: Proceedings of Museums and the Web (MW) (2004)
2. Karaman, S., Bagdanov, A.D., Landucci, L., D'Amico, G., Ferracani, A., Pezzatini, D., Del Bimbo, A.: Personalized multimedia content delivery on an interactive table by passive observation of museum visitors. *Multimedia Tools Appl.* **75**(7), 3787–3811 (2016)
3. Seidenari, L., Baccchi, C., Uricchio, T., Ferracani, A., Bertini, M., Bimbo, A.D.: Deep artwork detection and retrieval for automatic context-aware audio guides. *ACM Trans. Multimedia Comput. Commun. Appl.* **13**(3s), 35:1–35:21 (2017)
4. Baraldi, L., Paci, F., Serra, G., Benini, L., Cucchiara, R.: Gesture recognition using wearable vision sensors to enhance visitors' museum experiences. *IEEE Sens. J.* **15**(5), 2705–2714 (2015)
5. Abowd, G.D., Atkeson, C.G., Hong, J., Long, S., Kooper, R., Pinkerton, M.: Cyberguide: a mobile context-aware tour guide. *Wireless Netw.* **3**(5), 421–433 (1997)
6. Wang, Y., Stash, N., Sambeek, R., Schuurmans, Y., Aroyo, L., Schreiber, G., Gorgels, P.: Cultivating personalized museum tours online and on-site. *Interdisc. Sci. Rev.* **34**(2–3), 139–153 (2009)
7. Zancanaro, M., Kuflik, T., Boger, Z., Goren-Bar, D., Goldwasser, D.: Analyzing museum visitors' behavior patterns. In: Conati, C., McCoy, K., Paliouras, G. (eds.) UM 2007. LNCS (LNAI), vol. 4511, pp. 238–246. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73078-1_27
8. Kuflik, T., Boger, Z., Zancanaro, M.: Analysis and prediction of museum visitors' behavioral pattern types. In: Krüger, A., Kuflik, T. (eds.) Ubiquitous Display Environments. Cognitive Technologies, pp. 161–176. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-27663-7_10
9. Eliseo, V., Martine, L.: *Ethnographie de l'exposition. Études et recherche*, Centre Georges Pompidou, Bibliothèque publique d'information (1991)
10. Keil, J., Pujol, L., Roussou, M., Engelke, T., Schmitt, M., Bockholt, U., Eleftheratou, S.: A digital look at physical museum exhibits: designing personalized stories with handheld augmented reality in museums. In: Proceedings of Digital Heritage International Congress (DigitalHeritage) (2013)
11. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR) (2016)

Automatic Image Cropping and Selection Using Saliency: An Application to Historical Manuscripts

Marcella Cornia^(✉), Stefano Pini, Lorenzo Baraldi, and Rita Cucchiara

University of Modena and Reggio Emilia, Modena, Italy
{marcella.cornia, stefano.pini, lorenzo.baraldi,
rita.cucchiara}@unimore.it

Abstract. Automatic image cropping techniques are particularly important to improve the visual quality of cropped images and can be applied to a wide range of applications such as photo-editing, image compression, and thumbnail selection. In this paper, we propose a saliency-based image cropping method which produces significant cropped images by only relying on the corresponding saliency maps. Experiments on standard image cropping datasets demonstrate the benefit of the proposed solution with respect to other cropping methods. Moreover, we present an image selection method that can be effectively applied to automatically select the most representative pages of historical manuscripts thus improving the navigation of historical digital libraries.

Keywords: Image cropping · Image selection · Saliency
Digital libraries

1 Introduction

Image cropping aims at extracting rectangular subregions of a given image with the aim of preserving most of its visual content and enhancing the visual quality of the cropped image [5, 6, 30]. A good image cropping algorithm can have several applications, from helping professional editors in the advertisement and publishing industry, to increasing the presentation quality in search engines and social networks, where it is often the case that variable sized images need to be previewed with thumbnails of given size. In the case of collections of images, the combination of frame selection and image cropping techniques can be exploited to generate high quality thumbnails representing the entire collection. The same line of thinking can be extended, of course, to the case of selecting appropriate thumbnail for a video.

Multimedia digital libraries, which contain collections of images and videos [2, 4, 13], are for sure a valuable application domain of image cropping and selection techniques. Motivated by these considerations, in this paper we devise a cropping technique based on saliency prediction. In fact, visual saliency prediction is

the task of predicting the most important regions of an image by identifying those regions which most likely attract human gazes at the first glance [10–12]. By relying on this information, we propose a simple and effective image cropping solution which returns cropped regions with the most important visual content of their corresponding original images. To validate the effectiveness of the proposed cropping technique, we assess its performance on standard image cropping datasets by comparing to state of the art methods.

Moreover, we propose an image selection method which exploits the ability of our cropping solution of finding the most important regions of images. In particular, to validate our solution in real-world scenarios, we apply it to the selection of the most representative pages of historical manuscripts. In this way, the selected pages can be used as an effective preview of each manuscript thus improving the navigation of historical digital libraries.

Overall, the paper is organized as follows: Sect. 2 presents the main related image cropping methods and briefly reviews the thumbnail selection literature, Sect. 3 introduces the proposed saliency-based cropping technique, while the corresponding experimental results are reported in Sect. 4. Finally, the automatic page selection of historical manuscripts is presented in Sect. 5.

2 Related Work

In this section, we start from reviewing the literature related to the automatic image cropping task. Also, we briefly describe some recent works addressing the thumbnail selection problem.

2.1 Image Cropping

Existing image cropping methods can be categorized into two main categories: *attention-based* and *aesthetics-based* methods. The first ones aim at finding the most visually salient regions in the original images, while the second ones accomplish the cropping task mainly by analyzing the attractiveness of the cropped image with the help of a quality classifier.

Attention-based approaches exploit visual saliency models or salient object detectors to identify the crop windows that more attract human attention [5, 24, 26, 27]. Some other hybrid methods employ a face detector to locate the regions of interest [32] or directly fit a saliency map from visually pleasurable photos taken by professional photographers [23]. Instead of using saliency, pixel importances can be also estimated using their objectness [9], or empirically defined energy functions [1, 21].

On the other hand, aesthetics-based methods leverage on photo quality assessment studies [3, 15, 28] using certain objective aspects of images, such as low level image features and empirical photographic composition rules. In particular, Nishiyama *et al.* [22] built a quality classifier using low level image features such as color histogram and Fourier coefficient from which they selected the cropped region with the highest quality score. Chen *et al.* [8] presented a method to

learn the spatial correlation distributions of two arbitrary patches in an image for generating an omni-context prior which serve as rules to guide the composition of professional photos. Zhang *et al.* [31], instead, proposed a probabilistic model based on a region adjacency graph to transfer aesthetic features from the training photo onto the cropped ones.

More recently, Yan *et al.* [30] proposed several features that accounts the removal of distracting content and the enhancement of overall composition. The influence of these features on crop solutions was learned from a training set of image pairs, before and after cropping by expert photographers. Other works, instead, exploit a RankSVM [6], working with features coming from the AlexNet model [16], or an aesthetics-aware deep ranking network [7] to classify each candidate window. Finally, Li *et al.* [6] formulated the automatic image cropping problem as a sequential decision-making process, and proposed an Aesthetics Aware Reinforcement Learning (A2-RL) model to solve this problem.

2.2 Thumbnail Selection

The thumbnail selection problem has been widely addressed especially in the video domain, in which a frame that is visually representative of the video is selected and used as a representation of the video itself. In our case, instead, we want to find the most significant image from a collection of images (*i.e.* the pages of an historical manuscript), which somehow it can be considered as a related problem to the video thumbnail selection.

Most conventional methods for video thumbnail selection have focused on learning visual representativeness purely from visual content [14,20], while more recent researches have addressed this problem as the selection of query-dependent thumbnails to supply specific thumbnails for different queries.

Liu *et al.* [18] proposed a reinforcement algorithm to rank the frames in each video, while a relevance model was employed to calculate the similarity between the video frames and the query keywords. Wang *et al.* [29] introduced a multiple instance learning approach to localize the tags into video shots and to select query-dependent thumbnail according to the tags.

In [19], instead, a deep visual-semantic embedding was trained to retrieve query-dependent video thumbnails. In particular, this method employs a deeply-learned model to directly compute the similarity between the query and video thumbnails by mapping them into a common latent semantic space.

3 Automatic Image Cropping

We tackle the image cropping task as that of finding a rectangular region \mathcal{R} inside the given image \mathcal{I} with maximum saliency. Comparing to previous methods which maximized a function of the saliency inside \mathcal{R} , they all used other functions, such as the difference of saliency in \mathcal{R} and outside \mathcal{R} , or the difference between the mean saliency value in \mathcal{R} and the mean saliency value outside \mathcal{R} . We experimentally

validated that when using state of the art saliency predictors, our choice, although simple, provides better results than more fancy objective functions.

Formally, being \mathbf{x} a pixel of the input image and $S(\mathbf{x})$ its saliency value, predicted by a saliency model, we aim at finding:

$$\max_{\mathcal{R}} \left(\int_{\mathbf{x} \in \mathcal{R}} S(\mathbf{x}) - \int_{\mathbf{x} \in \mathcal{I} \setminus \mathcal{R}} S(\mathbf{x}) \right) \quad (1)$$

This objective boils down to finding the minimum bounding box of all salient pixels, and taking all regions \mathcal{R} which contains the minimum bounding box. Since taking regions larger than the minimum bounding box would amount to having non salient pixels in \mathcal{R} , we take \mathcal{R} as the minimum bounding box of salient pixels.

Regarding the saliency map, we compute it for every image by using the saliency method proposed in [12] which currently is the state of the art method in the saliency prediction task. In particular, starting from a classical convolutional neural network, it iteratively refines saliency predictions by incorporating an attentive mechanism. Also, it is able to reproduce the center bias present in human eye fixations by exploiting a set of prior maps directly learned from data. Overall, the performance achieved by the selected saliency method allows us to rely on saliency maps that effectively reproduce the human attention on natural images.

4 Experimental Evaluation

In this section, we briefly describe datasets and metrics used to evaluate our solution and provide quantitative and qualitative comparisons with other image cropping methods.

4.1 Datasets

To validate the effectiveness of visual saliency in the automatic image cropping task, we perform experiments on two different publicly available datasets.

The Flickr-Cropping dataset [6] is composed of 1,743 images, each of them associated to ground-truth cropping parameters. Images are divided in training and test sets, respectively composed of 1,395 and 348 images. Our method is not trainable, but we perform experiments on test images only for a fair comparison with other methods.

The CUHK Image Cropping dataset [30] contains the cropping parameters for 950 images that were manually cropped by an experienced photographer. Images are provided with cropping annotations of three different photographers. In our experiments, we evaluate the performance of our saliency-based cropping method with respect to all three different annotations.

4.2 Metrics

Two different metrics are usually used to determine the accuracy of the automatic image cropping algorithms: the Intersection over Union (commonly abbreviated as IoU) and the Boundary Displacement Error (BDE).

The Intersection over Union is an evaluation metric used to evaluate the overlapping between two bounding boxes. Technically, it is defined as

$$\text{IoU} = \frac{1}{N} \sum_i^N \frac{GT_i \cap P_i}{GT_i \cup P_i} \quad (2)$$

where N is the number of samples, GT_i is the area of the i th ground-truth bounding box and P_i is the area of the i th predicted bounding box.

The Boundary Displacement Error measures the distance between the sides of the ground-truth bounding box and the predicted one. For convenience, the values are normalized with respect to the size of the image. Mathematically, the metric is defined as

$$\text{BDE} = \frac{1}{4} \frac{1}{N} \sum_i^N \left(\frac{|x_1^{GT_i} - x_1^{P_i}|}{w_i} + \frac{|y_1^{GT_i} - y_1^{P_i}|}{h_i} + \frac{|x_2^{GT_i} - x_2^{P_i}|}{w_i} + \frac{|y_2^{GT_i} - y_2^{P_i}|}{h_i} \right) \quad (3)$$

where N is the number of samples, (x_1, y_1) is the top left edge of the bounding box, (x_2, y_2) is the bottom right edge of the bounding box, w_i and h_i are respectively width and height of the image, GT_i is the i th ground-truth bounding box, and P_i is the i th predicted bounding box.

4.3 Results

We compare our solution with other automatic image cropping methods. For the Flickr-Cropping dataset, we perform comparisons with the most competitive saliency-based baseline presented in [6] (eDN), the RankSVM+DeCAF₇ model [6],

Table 1. Experimental results on the Flickr-Cropping [6] dataset. First, second and third best scores on each metric are respectively highlighted in red, green and blue colors.

Method	Avg IoU	Avg BDE
eDN [6]	0.4857	0.1372
RankSVM+DeCAF ₇ [6]	0.6019	0.1060
VFN [7]	0.6744	0.0872
A2-RL [17]	0.6564	0.0914
Saliency density	0.6193	0.0997
VGG activations	0.6004	0.1088
Ours	0.6589	0.0892

the View Finding Network (VFN) proposed in [7] and the Aesthetics Aware Reinforcement Learning (A2-RL) model [17]. For the CUHK Image Cropping dataset, instead, the comparison methods are the change-based image cropping architecture presented in [30] (LearnChange) and the VFN and A2-RL models.

Moreover, for both datasets, we compare our results with two variations of our model which we call **Saliency Density** and **VGG Activations**. The first one aims at maximizing the difference of the averaged saliency between the selected bounding box and the outer region of the image. For simplicity, we set the size of search window to each scale among $[0.75, 0.80, \dots, 0.95]$ of the original image and slide the search window over a 10×10 uniform grid. The **VGG Activations** is, instead, the proposed image cropping method where the saliency maps are replaced with the activations of the last convolutional layer of the VGG-16 network [25]. In particular, since the last convolutional layer has 512 filters, we select for each image the activation map having the maximum sum.

Table 1 shows the results on the Flickr-Cropping dataset. As it can be seen, our solution obtains the second best scores on both IoU and BDE metrics and achieves better results with respect to both our baselines. Table 2, instead, reports the results on the three different annotations of the CUHK Image

Table 2. Experimental results on three different annotations of the CUHK Image Cropping [30] dataset. First, second and third best scores on each metric are respectively highlighted in red, green and blue colors.

Annotation	Method	Avg IoU	Avg BDE
1	LearnChange [30]	0.7487	0.0667
	VFN [7]	0.7847	0.0581
	A2-RL [17]	0.7934	0.0545
	Saliency density	0.6345	0.0971
	VGG activations	0.7788	0.0574
	Ours	0.8017	0.0500
2	LearnChange [30]	0.7288	0.0720
	VFN [7]	0.7763	0.0614
	A2-RL [17]	0.7911	0.0554
	Saliency density	0.6053	0.1075
	VGG activations	0.7648	0.0624
	Ours	0.7711	0.0594
3	LearnChange [30]	0.7322	0.0719
	VFN [7]	0.7602	0.0653
	A2-RL [17]	0.7826	0.0551
	Saliency density	0.6153	0.1040
	VGG activations	0.7612	0.0618
	Ours	0.7675	0.0599

Cropping dataset. In this case, our method achieves the best results on the first annotation on both metrics, while, on the other two annotations, it obtains the second or the third best scores. Despite the proposed solution is much simpler than the other comparison methods, the results achieved by our method on both considered datasets are very close to the best ones, thus confirming the effectiveness of the proposed strategy. Finally, some qualitative results with the corresponding saliency maps are presented in Fig. 1.

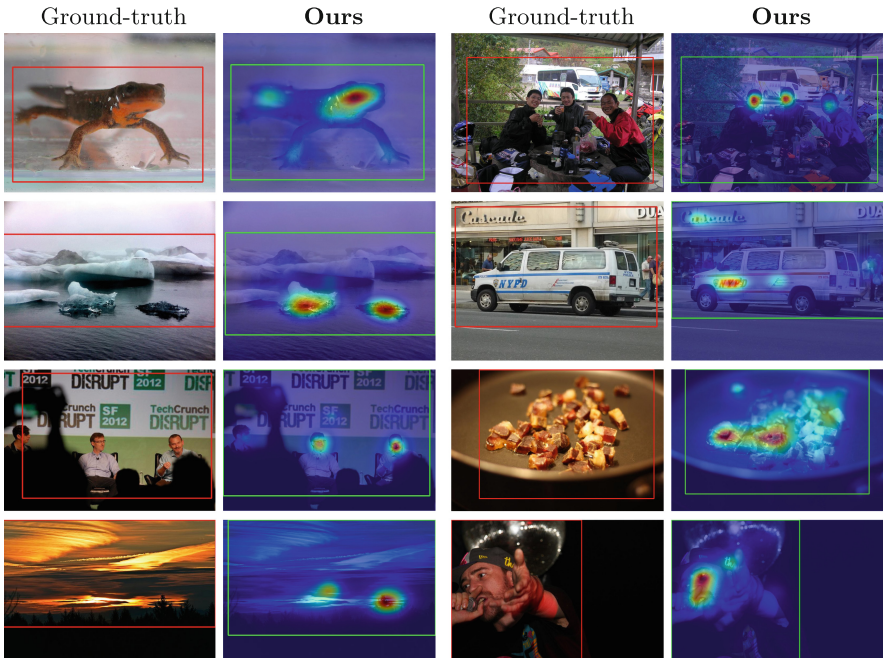


Fig. 1. Cropping results on sample images from the Flickr-Cropping dataset [6].

5 Automatic Page Selection of Historical Manuscripts

To validate our architecture in a real-world scenario, we apply it to find the best pages that represent historical manuscripts. This type of books usually have anonymous covers that does not represent its content, like plain colours or little artworks. Therefore, we develop a method to extract the most illustrative pages from every manuscript in order to use them as the preview of the book itself. Using this system, the navigation of historical digital libraries can be improved: users will be able to visually identify the content of a book watching its most representative images, without the need of opening it or read its summary.

In this case, the proposed image cropping method is not the output of the system, but it is used to find the most interesting pages of every manuscript.

In particular, the saliency map is calculated for every page of the book using the saliency model reported in [12]. After extracting all saliency maps,



Fig. 2. Example results of the page selection method on historical manuscripts. For each manuscript, the figure shows a list of some sample pages and the three pages selected by our method. As it can be seen, the selected pages contains representative visual contents and can be successfully used as a preview of the considered manuscript.

the method proposed in Sect. 3 is used to find the minimum crop that contains all the pixels with a saliency value higher than a threshold t (in our experiments $t = 128$). Then, a density score is calculated as the average value of saliency inside the bounding box divided by the average value of saliency outside the bounding box. In particular, it is formulated as

$$DS = \frac{\frac{1}{K} \sum_{i,j} s(i,j)}{\frac{1}{w \cdot h - K} \sum_{l,m} s(l,m)} \quad (4)$$

where K is the number of pixels inside the bounding box, (i, j) and (l, m) are respectively the coordinates of the pixels inside and outside the bounding box, while w and h are width and height of the image.

An high density score corresponds to an image where most of the saliency is restricted to a small area, therefore it contains a tiny region of high interest with respect to the rest of the image. On the contrary, a low density score corresponds to an image with a spread saliency map, therefore the image does not contain a valuable detail. Finally, the M images with the higher density score are selected as the most representative of the document.

Note that the method does not require training and it is applicable to any type of book, but it performs better with illustrated books. In our experiments, we decide to select entire images in place of image crops since we consider the full pages more suitable to be a summary of the whole manuscript, but it would be also possible to extract some particular details.

To validate our proposal, we apply the proposed automatic page selection method to a set of digitized historical manuscripts belonging to the Estense Library collection of Modena¹. Some notable results are shown in Fig. 2. As it can be seen, the selected pages contain representative visual contents of the corresponding manuscript and they can be used as a significant preview of the manuscript itself.

6 Conclusions

In this work, we presented a saliency-based image cropping method which, by selecting the minimum bounding box that contains all salient pixels, achieves promising results on different image cropping datasets. Moreover, we applied our solution to the image selection problem. In particular, to validate the effectiveness in real-world scenarios, we introduced a page selection method which identifies the most representative pages of an historical manuscript. Qualitative results demonstrated that our idea improves the navigation of historical digital libraries by automatic generating significant book previews.

Acknowledgment. We gratefully acknowledge the Estense Gallery of Modena for the availability of the digitized historical manuscripts used in this work. We also acknowledge the CINECA award under the ISCRA initiative, for the availability of high performance computing resources and support.

¹ <http://bibliotecaestense.beniculturali.it>.

References

1. Avidan, S., Shamir, A.: Seam carving for content-aware image resizing. *ACM Trans. Graph.* **26**(3), 10 (2007)
2. Balducci, F., Grana, C.: Affective classification of gaming activities coming from RPG gaming sessions. In: Tian, F., Gatzidis, C., El Rhalibi, A., Tang, W., Charles, F. (eds.) *Edutainment 2017. LNCS*, vol. 10345, pp. 93–100. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65849-0_11
3. Bhattacharya, S., Sukthankar, R., Shah, M.: A framework for photo-quality assessment and enhancement based on visual aesthetics. In: *ACM International Conference on Multimedia* (2010)
4. Bolelli, F.: Indexing of historical document images: ad hoc dewarping technique for handwritten text. In: Grana, C., Baraldi, L. (eds.) *IRCDL 2017. CCIS*, vol. 733, pp. 45–55. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68130-6_4
5. Chen, J., Bai, G., Liang, S., Li, Z.: Automatic image cropping: a computational complexity study. In: *IEEE International Conference on Computer Vision and Pattern Recognition* (2016)
6. Chen, Y.L., Huang, T.W., Chang, K.H., Tsai, Y.C., Chen, H.T., Chen, B.Y.: Quantitative analysis of automatic image cropping algorithms: a dataset and comparative study. In: *Winter Conference on Applications of Computer Vision* (2017)
7. Chen, Y.L., Klopp, J., Sun, M., Chien, S.Y., Ma, K.L.: Learning to compose with professional photographs on the web. *arXiv preprint arXiv:1702.00503* (2017)
8. Cheng, B., Ni, B., Yan, S., Tian, Q.: Learning to photograph. In: *ACM International Conference on Multimedia* (2010)
9. Ciocca, G., Cusano, C., Gasparini, F., Schettini, R.: Self-adaptive image cropping for small displays. *IEEE Trans. Consum. Electron.* **53**(4), 1622–1627 (2007)
10. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: A deep multi-level network for saliency prediction. In: *International Conference on Pattern Recognition* (2016)
11. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Multi-level net: a visual saliency prediction model. In: *European Conference on Computer Vision Workshops* (2016)
12. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Predicting human eye fixations via an LSTM-based saliency attentive model. *arXiv preprint arXiv:1611.09571* (2017)
13. Cucchiara, R., Grana, C., Prati, A.: Semantic transcoding for live video server. In: *ACM International Conference on Multimedia* (2002)
14. Kang, H.W., Hua, X.S.: To learn representativeness of video frames. In: *ACM International Conference on Multimedia* (2005)
15. Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: *IEEE International Conference on Computer Vision and Pattern Recognition* (2006)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012)
17. Li, D., Wu, H., Zhang, J., Huang, K.: A2-RL: aesthetics aware reinforcement learning for automatic image cropping. *arXiv preprint arXiv:1709.04595* (2017)
18. Liu, C., Huang, Q., Jiang, S.: Query sensitive dynamic web video thumbnail generation. In: *IEEE International Conference on Image Processing* (2011)
19. Liu, W., Mei, T., Zhang, Y., Che, C., Luo, J.: Multi-task deep visual-semantic embedding for video thumbnail selection. In: *IEEE International Conference on Computer Vision and Pattern Recognition* (2015)

20. Luo, J., Papin, C., Costello, K.: Towards extracting semantically meaningful key frames from personal video clips: from humans to computers. *IEEE Trans. Circ. Syst. Video Technol.* **19**(2), 289–301 (2009)
21. Ma, M., Guo, J.K.: Automatic image cropping for mobile device with built-in camera. In: *Consumer Communications and Networking Conference* (2004)
22. Nishiyama, M., Okabe, T., Sato, Y., Sato, I.: Sensation-based photo cropping. In: *ACM International Conference on Multimedia* (2009)
23. Park, J., Lee, J.Y., Tai, Y.W., Kweon, I.S.: Modeling photo composition and its application to photo re-arrangement. In: *IEEE International Conference on Image Processing* (2012)
24. Santella, A., Agrawala, M., DeCarlo, D., Salesin, D., Cohen, M.: Gaze-based interaction for semi-automatic photo cropping. In: *SIGCHI Conference on Human Factors in Computing Systems* (2006)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014)
26. Stentiford, F.: Attention based auto image cropping. In: *Workshop on Computational Attention and Applications, ICVS* (2007)
27. Suh, B., Ling, H., Bederson, B.B., Jacobs, D.W.: Automatic thumbnail cropping and its effectiveness. In: *ACM Symposium on User Interface Software and Technology* (2003)
28. Tang, X., Luo, W., Wang, X.: Content-based photo quality assessment. *IEEE Trans. Multimed.* **15**(8), 1930–1943 (2013)
29. Wang, M., Hong, R., Li, G., Zha, Z.J., Yan, S., Chua, T.S.: Event driven web video summarization by tag localization and key-shot identification. *IEEE Trans. Multimed.* **14**(4), 975–985 (2012)
30. Yan, J., Lin, S., Bing Kang, S., Tang, X.: Learning the change for automatic image cropping. In: *IEEE International Conference on Computer Vision and Pattern Recognition* (2013)
31. Zhang, L., Song, M., Zhao, Q., Liu, X., Bu, J., Chen, C.: Probabilistic graphlet transfer for photo cropping. *IEEE Trans. Image Process.* **22**(2), 802–815 (2013)
32. Zhang, M., Zhang, L., Sun, Y., Feng, L., Ma, W.: Auto cropping for digital photographs. In: *ICME* (2005)

Bidirectional LSTM Recurrent Neural Network for Keyphrase Extraction

Marco Basaldella^(✉), Elisa Antolli, Giuseppe Serra, and Carlo Tasso

Artificial Intelligence Laboratory, Department of Mathematics, Computer Science,
and Physics, University of Udine, Udine, Italy

{marco.basaldella, giuseppe.serra, carlo.tasso}@uniud.it,
antolli.elisa@spes.uniud.it

Abstract. To achieve state-of-the-art performance, keyphrase extraction systems rely on domain-specific knowledge and sophisticated features. In this paper, we propose a neural network architecture based on a Bidirectional Long Short-Term Memory Recurrent Neural Network that is able to detect the main topics on the input documents without the need of defining new hand-crafted features. A preliminary experimental evaluation on the well-known INSPEC dataset confirms the effectiveness of the proposed solution.

1 Introduction

Keyphrases (herein KPs) are phrases that “capture the main topic discussed on a given document” [31]. More specifically, KPs are phrases typically one to five words long that appear verbatim in a document, and can be used to briefly summarize its content.

The task of finding such KPs is called Automatic Keyphrase Extraction (herein AKE). It has received a lot of attention in the last two decades [11] and recently it has been successfully used in many Natural Language Processing (hence NLP) tasks, such as text summarization [34], document clustering [10], or non-NLP tasks such as social network analysis [23] or user modeling [24]. Automatic Keyphrase Extraction approaches have been also applied in Information Retrieval of relevant documents in digital document archives which can contain heterogeneous types of items, such as books articles, papers etc. [15].

The first approaches to solve Automatic Keyphrase Extraction were based on supervised machine learning (herein ML) algorithms, like Naive Bayes [32] or C4.5 decision trees [31]. Since then, several researchers explored different ML techniques such as Multilayer Perceptrons [2, 19], Support Vector Machines [19], Logistic Regression [2, 9], and Bagging [14]. Since no algorithm stands out as the “best” ML algorithm, often authors test many techniques in a single experiment, and then they choose as best ML algorithm the best performing one [2, 9] and/or even the least computationally expensive one [19].

M. Basaldella and E. Antolli—Equally Contributed.

However, AKE algorithms based on unsupervised approaches have been developed over the years as well. For example, Tomokiyo *et al.* [30] proposed to use a language model approach to extract KPs, and Mihalcea *et al.* [21] presented a graph-based ranking algorithm to find keyphrases. Nevertheless, supervised approaches have been the best performing ones in challenges: for example, [19], a supervised approach, was the best performing algorithm in the SEMEVAL 2010 Keyphrase Extraction Task [16].

In the last years, most attention is devoted to the *features* used in these supervised algorithms. The numbers of features used can range from just two [32] to more than 20 [9]. These features can be divided in categories identified with different kinds of knowledge they encode into the model:

- *statistical knowledge*: number of appearances of the KP in the document, TF-IDF, number of sentences containing the KP, etc.;
- *positional knowledge*: position of the first occurrence of the KP in the document, position of the last occurrence, appearance in the title, appearance in specific sections (abstract, conclusions), etc.;
- *linguistic knowledge*: part-of-speech tags of the KP [14], anaphoras pointing to the KP [2], etc.;
- *external knowledge*: presence of the KP as a page on Wikipedia [6] or in specialized domain ontologies [19], etc.

However, given the wide variety of lexical, linguistic and semantic aspects that can contribute to define a keyphrase, it difficult to design hand-crafted feature, and even the best performing algorithms hardly reach F1-Scores of 50% on the most common evaluation sets [14,16]. For this reason, AKE is still far from being a solved problem in the NLP community.

In recent years, Deep Learning techniques have shown impressive results in many Natural Language Processing tasks, e.g., Named Entity Recognition, Automatic Summarization, Question Answering, and so on [18,25,27,29]. In Named Entity Recognition, for example, researchers have proposed several Neural Network Architectures

To best of our knowledge, only recently some first attempts to address AKE task with Deep Learning techniques, has been presented [20,33]. In [33], the authors present an approach based on Recurrent Neural Networks, specifically designed for a particular domain, i.e., Twitter data. On the other hand, in [20] the authors use more datasets to evaluate their RNN for keyphrase extraction, and they propose a study of the keyphrases *generated* by their network as well.

In this paper, we present a Deep Learning architecture for AKE. In particular, we investigate an approach based on based on Bidirectional Long Short-Term Memory RNN (hence Bi-LSTM), which is able to exploit previous and future context of a given word. Our system, since it does not require specific features carefully optimized for a specific domain, can be applied to a wide range of scenarios. To evaluate the proposed method, we conduct experiments on the well-known INSPEC dataset [14]. The experimental result showed the proposed solution performs significantly better than competitive methods.

2 Proposed Approach

To extract KPs we implemented the following steps, as presented in Fig. 1. First, we split the document into sentences, and then we tokenize the sentences in words using NLTK [3]. Then, we associate a word embedding representation that maps each input word into a continuous vector representation. Finally, we feed our word embeddings into a Bi-LSTM units, which it can effectively deal with the variable lengths of sentences and it is able to analyze word features and their context (for example, distant relation between words). The Bi-LSTM is connected to a fully connected hidden layer, which in turn is connected to a softmax output layer with three neurons for each word. Between the Bi-LSTM layer and the hidden layer, and between the hidden layer and the output layer, we use dropout [28] to prevent overfitting.

As in the techniques used for Named Entity Recognition, the three neurons are mapped to three possible output classes: NO_KP, BEGIN_KP, INSIDE_KP, which respectively mark tokens that are *not* keyphrases, the *first* token of a keyphrase, and the other tokens of a keyphrase.

For example, if our input sentence is “*We train a neural network using Keras*”, and the keyphrases in that sentence are “*neural network*” and “*Keras*”, the tokens’ classes will be We/NO_KP train/NO_KP a/NO_KP neural/BEGIN_KP network/INSIDE_KP using/NO_KP Keras/BEGIN_KP’.

2.1 Word Embeddings

The input layer of our model is a vector representation of the individual words contained in input document. Several recent studies [5, 22] showed that such representations, called word embeddings, are able to represent the semantics of words better than an “one hot” encoding word representation, when trained on large corpus. However, the datasets for AKE are relatively small, therefore

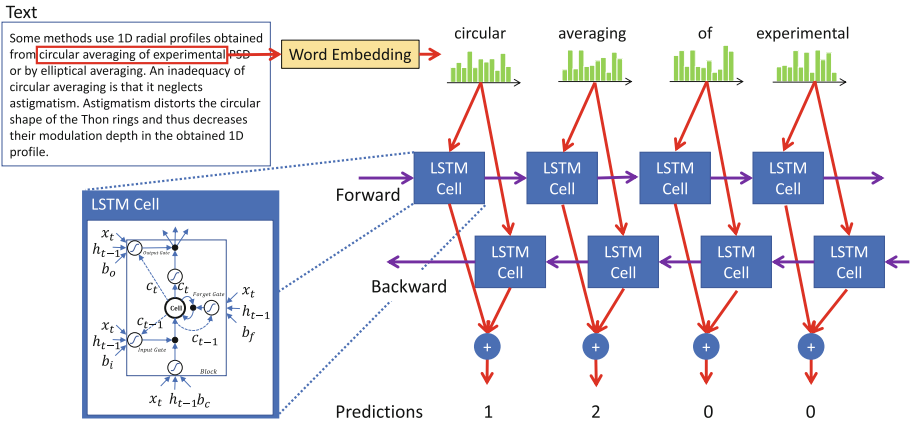


Fig. 1. Overview of the proposed system.

it is difficult to train word embeddings to capture the word semantics. Hence, we adopt Stanford’s GloVe Embeddings, which are trained on 6 billion words extracted from Wikipedia and Web texts [26].

2.2 Model Architecture

Let $\{x_1, \dots, x_n\}$ the word embeddings representing the input tokens, a Recurrent Neural Network (hence RNN) computes the output vector y_t of each token x_t by iterating the following equations from $t = 1$ to n :

$$h_t = H(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (1)$$

$$y_t = W_{hy}h_t + b_y \quad (2)$$

where h_t is the hidden vector sequence, W denotes weight matrices (for example W_{xh} is the matrix of the weights connecting the input layer and the hidden layer), b denotes bias vectors, and H is activation function of the hidden layer. Equation 1 represents the connection between the previous and the current hidden states, thus RNNs can make use of previous context.

In practice however, the RNN is not able to use effectively the all input history due to the *vanishing gradient* problem [12]. Hence, a better solution to exploit long range context is the Long Short-Term Memory (LSTM) architecture [13]. The LSTM is conceptually defined like an RNN, but hidden layer updates are replaced by specific units called memory cells. Specifically, a LSTM is implemented by the following functions [7]:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (4)$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (5)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (6)$$

$$h_t = o_t \tanh(c_t) \quad (7)$$

where σ is the logistic sigmoid function, i , f , o , and c are the input gate, forget gate, output gate and cell activation vectors, and all b are learned biases.

Another shortcoming of RNNs is that they consider only previous context, but in AKE we want to exploit future context as well. For example, consider the phrase “John Doe is a lawyer; he likes fast cars”. When we first encounter “John Doe” in the phrase, we still don’t know whether he’s going to be an important entity; then, we find the word “lawyer” and the pronoun “he”, which clearly refer to him, stressing his importance in the context. “Lawyer” and “he” are called *anaphoras* and the technique to find this contextual information is called *anaphora resolution*, which has been exploited to perform keyphrase extraction in [2].

In order to use future context, in our approach we adopt a Bidirectional LSTM network [8]. In fact, with this architecture we are able to make use of both past context and future context of a specific word. It consists of two separate hidden layers; it first computes the forward hidden sequence \vec{h}_t ; then, it computes

the backward hidden sequence \overleftarrow{h}_t ; finally, it combines \overrightarrow{h}_t and \overleftarrow{h}_t to generate the output y_t . Let the hidden states h be LSTM blocks, a Bi-LSTM is implemented by the following functions:

$$\overrightarrow{h}_t = H(W_{x\overrightarrow{h}}x_t + W_{\overrightarrow{h}\overrightarrow{h}}\overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}) \quad (8)$$

$$\overleftarrow{h}_t = H(W_{x\overleftarrow{h}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}) \quad (9)$$

$$y_t = W_{\overrightarrow{h}y}\overrightarrow{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y \quad (10)$$

3 Experimental Results

We present experiments on a well-known keyphrase extraction dataset: the INSPEC dataset [14]. It is composed by 2000 abstract papers in English extracted from journal papers from the disciplines Computer and Control, Information Technology. It consists of 1000 documents for training, 500 for validation and the remaining 500 for testing. We choose this dataset since it's well known in the AKE community, so there are many other available results to compare with; moreover, is much bigger than the dataset used in the SEMEVAL 2010 [16] competition, which contains only 144 documents for training, 40 for validation, and 100 for testing.

In order to implement our approach, we used Keras with Theano [1] as back end, which in turn allowed us to use CUDA to train our networks using a GPU. Experiments are run on a GeForce GTX Titan X Pascal GPU. The network is trained to minimize the Crossentropy loss. We train our network using the Root Mean Square Propagation optimization algorithm [17] and batch size 32. After trying different configurations for the network, we obtained the best results with a size of 150 neurons for the Bi-LSTM layer, 150 neurons for the hidden dense layer, and a value of 0.25 for the dropout layers in between.

To test the impact of word embeddings, we perform experiments with the pre-trained Stanford's GloVe Embeddings using all the word embedding sizes available, i.e., 50, 100, 200 and 300. The training of the network takes about 30 s to perform a full epoch with all the GloVe Embeddings. To stop the training, we used Keras' own embedded early stopping rule, which halts training when the training loss does not decrease for two consecutive epochs. The number of epochs requested to converge in all the four settings is displayed in Table 1, along with precision, recall and F1-score obtained by our system when trained using different sizes of the word embeddings. We can note that the best results are obtained with embedding size of 100; however, the embedding sizes of 200 and 300 obtain a very close result in term of F1-Score. The scores seem to show an interesting pattern: in fact, looking at the results, we see that the precision increases with embedding size, while recall decreases from size 100 onwards.

Table 2 compares the performances in term of precision, recall, and F-score our approach with other competitive systems, based both on supervised and unsupervised machine learning techniques. The first three systems are the ones

Table 1. Performance with different vector sizes of the GloVe Word Embeddings: 50, 100, 200 and 300 (we called them GloVe-(SIZE), respectively).

Embedding	Size	Precision	Recall	F1-score	Epochs
GloVe-50	50	0.331	0.518	0.404	20
GloVe-100	100	0.340	0.578	0.428	14
GloVe-200	200	0.352	0.539	0.426	18
GloVe-300	300	0.364	0.500	0.421	8

Table 2. Comparison results on INSPEC dataset

Method	Precision	Recall	F1-score
Proposed approach	0.340	0.578	0.428
<i>n</i> -grams with tag [14]	0.252	0.517	0.339
NP chunking with tag [14]	0.297	0.372	0.330
Pattern with tag [14]	0.217	0.399	0.281
TopicRank [4]	0.348	0.404	0.352

presented in [14], with three different candidate keyphrase generation techniques: *n*-grams, Noun Phrase (NP) chunking, and patterns. The fourth system is TopicRank [4], a graph-based keyphrase extraction method that relies on a topical representation of the document. Our proposed solution achieves best performance in term of F1-score and Recall. Although TopicRank obtains best performance in precision, its recall results are significantly worse than the ones obtained by us; moreover, we have to stress that we’re able to obtain better precision when using an embedding size of 200 and 300, albeit with a slightly lower overall F1-Score. Finally, it’s worth noting that we perform better than the results presented in [20], which is to the best of our knowledge the only one DL AKE algorithm evaluated on the INSPEC dataset. In fact, we obtain a F1@10 score of 0.422, while the best F1@10 score obtained by [20] is 0.342.

4 Conclusion

In this work, we proposed a Deep Long-Short Term Memory Neural Network model to perform automatic keyphrase extraction, evaluating the proposed method on the INSPEC dataset. Since word representation is a crucial step for success, we perform experiments with different pre-trained word representations. We show that without requiring hand-crafted features, the proposed approach is highly effective and achieves better results with respect to other competitive methods. For the future, we plan to test additional network architectures and to evaluate our algorithms on more datasets, in order to demonstrate its robustness.

References

1. Al-Rfou, R., et al.: Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints abs/1605.02688, May 2016, <http://arxiv.org/abs/1605.02688>
2. Basaldella, M., Chiaradia, G., Tasso, C.: Evaluating anaphora and coreference resolution to improve automatic keyphrase extraction. In: Proceedings of International Conference on Computational Linguistics (2016)
3. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python, 1st edn. O'Reilly Media Inc., Sebastopol (2009)
4. Bougouin, A., Boudin, F., Daille, B.: Topicrank: graph-based topic ranking for keyphrase extraction. In: Proceedings of International Joint Conference on Natural Language Processing (2013)
5. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **12**, 2493–2537 (2011)
6. Degl'Innocenti, D., De Nart, D., Tasso, C.: A new multi-lingual knowledge-base approach to keyphrase extraction for the Italian language. In: Proceedings of International Conference on Knowledge Discovery and Information Retrieval (2014)
7. Gers, F.A., Schraudolph, N.N., Schmidhuber, J.: Learning precise timing with LSTM recurrent networks. *J. Mach. Learn. Res.* **3**, 115–143 (2002)
8. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **18**(5), 602–610 (2005)
9. Haddoud, M., Abdeddaim, S.: Accurate keyphrase extraction by discriminating overlapping phrases. *J. Inf. Sci.* **40**(4), 488–500 (2014)
10. Hammouda, K.M., Matute, D.N., Kamel, M.S.: CorePhrase: keyphrase extraction for document clustering. In: Perner, P., Imiya, A. (eds.) *MLDM 2005. LNCS (LNAI)*, vol. 3587, pp. 265–274. Springer, Heidelberg (2005). https://doi.org/10.1007/11510888_26
11. Hasan, K.S., Ng, V.: Automatic keyphrase extraction: a survey of the state of the art. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (2014)
12. Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J.: Gradient flow in recurrent nets: the difficulty of learning long-term dependencies (2001)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
14. Hulth, A.: Improved automatic keyword extraction given more linguistic knowledge. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (2003)
15. Jones, S., Staveley, M.S.: Phrasier: a system for interactive document retrieval using keyphrases. In: Proceedings of International ACM SIGIR Conference on Research and development in Information Retrieval (1999)
16. Kim, S.N., Medelyan, O., Kan, M.Y., Baldwin, T.: Semeval-2010 task 5: automatic keyphrase extraction from scientific articles. In: Proceedings of International Workshop on Semantic Evaluation (2010)
17. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings of International Conference on Learning Representations (2014)
18. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of International Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2016)

19. Lopez, P., Romary, L.: HUMB: automatic key term extraction from scientific articles in GROBID. In: Proceedings of International Workshop on Semantic Evaluation (2010)
20. Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., Chi, Y.: Deep keyphrase generation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, pp. 582–592. Association for Computational Linguistics (2017). <http://aclanthology.coli.uni-saarland.de/pdf/P/P17/P17-1054.pdf>
21. Mihalcea, R., Tarau, P.: Texttrank: bringing order into texts. In: Proceedings of Empirical Methods on Natural Language Processing (2004)
22. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
23. De Nart, D., Degl’Innocenti, D., Basaldella, M., Agosti, M., Tasso, C.: A content-based approach to social network analysis: a case study on research communities. In: Calvanese, D., De Nart, D., Tasso, C. (eds.) IRCDL 2015. CCIS, vol. 612, pp. 142–154. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41938-1_15
24. De Nart, D., Degl’Innocenti, D., Pavan, A., Basaldella, M., Tasso, C.: Modelling the User Modelling Community (and Other Communities as Well). In: Ricci, F., Bontcheva, K., Conlan, O., Lawless, S. (eds.) UMAP 2015. LNCS, vol. 9146, pp. 357–363. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20267-9_31
25. Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., Ward, R.: Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(4), 694–707 (2016)
26. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of Empirical Methods on Natural Language Processing (2014)
27. Rush, A.M., Chopra, S., Weston, J.: A neural attention model for abstractive sentence summarization. arXiv preprint [arXiv:1509.00685](https://arxiv.org/abs/1509.00685) (2015)
28. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)
29. Tan, M., Xiang, B., Zhou, B.: LSTM-based deep learning models for non-factoid answer selection. CoRR abs/1511.04108 (2015). <http://arxiv.org/abs/1511.04108>
30. Tomokiyo, T., Hurst, M.: A language model approach to keyphrase extraction. In: Proceedings of Workshop on Multiword Expressions: Analysis, Acquisition and Treatment (2003)
31. Turney, P.D.: Learning algorithms for keyphrase extraction. *Inf. Retrieval.* **2**(4), 303–336 (2000)
32. Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G.: KEA: practical automatic keyphrase extraction. In: Proceedings of ACM Conference on Digital Libraries, pp. 254–255 (1999)
33. Zhang, Q., Wang, Y., Gong, Y., Huang, X.: Keyphrase extraction using deep recurrent neural networks on Twitter. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (2016)
34. Zhang, Y., Zincir-Heywood, N., Milios, E.: World Wide Web site summarization. *Web Intell. Agent Syst.* **2**(1), 39–53 (2004)

Models and Applications

Eliciting the Ancient Geography from a Digital Library of Latin Texts

Maurizio Lana¹ and Timothy Tambassi²(✉)

¹ Dipartimento di Studi Umanistici, Università del Piemonte Orientale,
Piazza Roma 36, 13100 Vercelli, Italy
maurizio.lana@uniupo.it

² ICUB, University of Bucharest, 1 Dimitrie Brandza St.,
060102 Bucharest, Romania
timothy.tambassi@gmail.com

Abstract. *Geolat – Geography for Latin Literature* is a research project, aimed at making accessible a digital library containing the works of Latin literature (from its origins in 240 BCE to the end of the Roman Empire in 476 CE) through a query interface of geographic/cartographic type representing the geographic knowledge expressed in the Latin texts themselves. A core activity of the project has been the development of the ontology *GO!*, which describes the geographical knowledge contained in the texts of the library. The ontologically annotated texts will allow for a variety of scientifically relevant uses, apart from the geo-based browsing: for example the production of digital and printed critical editions. The project is under development at Dipartimento di Studi Umanistici of Università del Piemonte Orientale, and financially supported by Fondazione Compagnia di San Paolo.

Keywords: Geography · Ontology · OWL · Web · Classical latin texts
Digital library

1 Geolat, GO! and the Plurality of Areas of Research

Geolat – Geography for Latin Literature is a research project, aimed at making accessible a digital library containing the works of Latin literature (from its origins in 240 BCE to the end of the Roman Empire in 476 CE) through a query interface of geographic/cartographic type, representing the geographic knowledge expressed in the Latin texts themselves¹. The key points and the most relevant aspects of Geolat project are various (LOD, crowdsourcing, Open Access, CC licenses, semantic annotation of geographical references, URIs for annotation of places, and others) but here we want to focus on its ontology.

In particular, the introduction of a specific geo-ontology represents a fundamental innovation compared with similar projects focused on the ancient world. Some questions arise, the main ones being:

- what does ontology mean in this context?

¹ M. Lana is specifically responsible for Sects. 1 and 5, and T. Tambassi for Sects. 2–4.

- what kinds of problems does it deal with?
- what are the main objectives and features of *GO!* - the geo-ontology of Geolat?

First of all, we can distinguish three different disciplinary areas which make up this specific geo-ontological domain:

- computer science,
- contemporary geography,
- ancient geography.

In the domain of computer science, ontology is a structure aimed to describe the categorical hierarchy of a specific domain, analysing its basic constituents (entities like objects, events, processes, etc.), the properties characterizing them and the relationships which correlate them - using a language (usually OWL) that is understood both by the machines and the humans. The resulting structured representation of knowledge allows to resolve conceptual or terminological inconsistencies and provides a lexical or taxonomic framework for the representation of knowledge [7, 14, 16].

From a geographical point of view, the aim of a geo-ontology is to analyse the mesoscopic world of geographical partitions in order to:

- establish whether and what kinds of geographical entities exist, their borders, their spatial representation (in maps, software, etc.), their mereological and topological relations, and their location;
- determinate how they can be defined and classified in an ontological system which gather them together;
- argue whether and how the geographic descriptions of reality emerging from common sense can be combined with descriptions derived from different scientific disciplines [2, 3, 5, 8, 9, 15, 17–19].

Finally, if one wants to investigate the geographical knowledge expressed in the Latin texts² (which is a component of the Roman culture) some specific problems closely interconnected, and sharing the vagueness of data and information available, have to be taken into account. They can be distinguished in topological, source and methodological problems. Topological problems have to do e.g. with: measurement and measurability of distances (and their different units of measurement), location of places and absolute vs relative distances/coordinates. Problems concerning documentation and sources, have to do e.g. with: lack of reliability and homogeneity of some data, disagreement among authors, difficulty or impossibility of autoptical confirms and isolation of properly geographical contents from the rest of the texts. The third kind of problems is strictly connected with the second ones and refers to methods and to the (multiplicity of) approaches to ancient geographical investigation which involved e.g. heterogeneity of aims, points of view, interpretations and perspectives (sometimes overlapped) through which the information was transmitted, processed and implemented, the importance of imagination (and mental maps), the necessity of folk

² Until now the Latin texts are offered with no translation because even if translations do exist for them, recent translations are protected by Intellectual Property Rights; and if they are free from IPR they are generally speaking too remote from today sensibility to be usable.

theorizing, in order to understand other's mind and ancient culture [1, 6, 11]. All of these problems are mentioned because they mark the distance between the two (today and ancient) geographies involved in *GO!*.

2 Steps for the Creation of GO!

Given the plurality of domain interests, the creation of GO! [10] imposed a division of work in different steps: a search and analysis of the existing ontologies which can or could be of interest for the geography in order to start to understand to which extent they could be reused for the scope of describing the geographical knowledge contained in the Latin classical texts; this was done through a critical review of contemporary geo-ontologies³, aimed to identify common classes⁴ and properties⁵, and also missing classes and properties needed to describe ancient geography - the scope was to establish if an ontology had to be built *ex novo* or if more simply classes and properties can be selected and imported from other existing ontologies, emphasizing in this way the specific contribution of Geolat ontology to the contemporary debate.

After that the analysis of Latin literature texts started, in order to identify geographical entities, classes, properties and relations; practically speaking it meant that around 15.000 pages of translated Latin texts were read (no problem with the the translation because not concepts but things were searched for), and everything *related to geography in broad sense* was highlighted: proper nouns (e.g. Rome); common nouns (mons, mare); names of populations; space/place indications: above, below, beyond, etcetera; all the verbs having any geography related nuance: build, move, settle, etcetera; properties and relations related to these entities or describing them.

Then all the relevant passages were re-read in Latin in order to check for possible translation problems and the highlighted entities were listed by type and author subsequently agreements and disagreements - among the Latin authors were analysed and highlighted, focusing on their basic distinctions; the study of the differences between ancient and contemporary geography, in terms of domains, presuppositions, representations and vagueness; the scope was that of understanding was could be expressed

³ A thorough description of this phase is Tambassi T.: Rethinking Geo-Ontologies from a Philosophical Point of View. Journal of Research and Didactics in Geography (J-Reading) 2(5), 51–62 (2016), <https://doi.org/10.4458/7800-04>. Geo-Ontologies can be broadly distinguished among geomatics, topological and geometrical ontologies: see e.g. OGC GeoSPARQL, Spatial Schema – ISO 19107, Spatial referencing by coordinates - ISO 19111; physical and natural ontologies: see e.g. NDH Ontology (USGS) and Hydro Ontology (Spanish GeoData); human ontologies: see e.g. FAO Geopolitical Ontology. Well known ontologies like DOLCE, CIDOC-CRM, FRBR were not used because they don't offer a sufficiently detailed characterization of the geographical knowledge.

⁴ “Classes are used to group individuals that have something in common in order to refer to them. [...] In modeling, classes are often used to denote the set of objects comprised by a concept of human thinking, like the concept person or the concept woman.” *OWL 2 Web Ontology Language Primer (Second Edition)*, <https://www.w3.org/2007/OWL/wiki/Primer>.

⁵ “In OWL 2, we denote ... relations as *properties*. Properties are further subdivided: *Object properties* relate objects to objects (like a person to their spouse), while *datatype properties* assign data values to objects (like an age to a person)” (*id.*).

using existing ontologies, and what not, and at the end of this phase it was discussed what could be reused from existing ontologies and what had to be created *ex novo*; this produced a reunification of these information in a geo-ontology for Latin literature, based on common sense classes, properties and relations, and folk conceptualizations. It allows to improve the usability of this ontology, making it more compatible with similar ontologies and conceptualizations.

An ulterior check against the Latin texts of the first phase was made to be sure that the conceptualization which was arising was effectively usable to describe the knowledge contained in those texts; and an additional similar check was made with other authors (minor historians – Eutropius, Velleius Paterculus, etc.; plus, some works from Cicero, Seneca, Vergil, Plautus, Catullus, Terentius) to be sure that relevant concepts didn't appear which were forgotten.

3 GO!: A Geo-Ontology for Latin Literature

The result of this work⁶ is *GO!*, a geo-ontology which provides a description of the geographical knowledge emerging from Latin literature and an inventory of classes and relation mainly focused towards semantically annotating Latin texts, identifying the places mentioned in these texts, and connecting them with their contemporary equivalents. The fundamental scopes of this ontology are essentially four: informativeness, completeness, reusability and accessibility (both for the scientific community and for general public). The main challenge of the project has been to put together all the different disciplinary areas - which include, at least, Computer Science, Contemporary Geography, Ancient (History and) Geography, Latin Literature, Ontology of Geography - that constitute the domain of this geo-ontology. Accordingly, the idea behind this ontology was that the study of common sense (geographical) conceptualizations - that is the body of knowledge that (ancient) people have about the surrounding geographic world - could constitute a fundamental infrastructure for the ontological representation and for the communication among different areas of research.

Also thinking of the reuse the ontology is built as a collection of four interconnected modules (expressed in OWL2) freely accessible, readable, usable at the following IRIs (now that the PURL services does no more accept new contents):

<https://w3id.org/geolit/ontologies/GO-TOP>
<https://w3id.org/geolit/ontologies/GO-PHY>
<https://w3id.org/geolit/ontologies/GO-HUM>
<https://w3id.org/geolit/ontologies/GO-FAR>

The modules are open to the use of all the interested people under a CC BY-NC-SA license; modification locally managed are discouraged because they create unnecessary

⁶ The group working to the ontology was made by a computer scientist, Diego Magro; a philosophy postdoc, Timothy Tambassi; a digital humanist, Maurizio Lana; plus a group of people who wrote and then revised the OWL ontology: Claudia Corcione, Paola De Caro, Silvia Naro and Marco Rovera. The group comprises also Gabriella Vanotti, ancient history; Cristina Meini, philosophy of language; Margherita Benzi, philosophy of science; and †Roberta Piastrri, Latin literature.

forks of the ontology; requests to the managing team are instead welcome. Graphical representations of the modules can be found at this address: <http://vowl.visualdataweb.org/webvowl/index.html#iri=https://w3id.org/geolit/ontologies/GO-TOP> (replace GO-TOP with GO-PHY, GO-HUM, GO-FAR for the graphical view of the other ontologies).

4 The Modules and the Modelling Choices

GO-TOP contains 21 classes, 38 object properties, 15 datatype properties, 4 individuals. It is the *top level ontology* which connects all the other modules and contains the most general elements that describe all the geographic entities included in GO!. In particular, all the most general classes and (object and data) properties belong to the GO-TOP and are used by the other three modules.

GO-PHY contains 127 classes, 3 individuals. It imports the GO-TOP module, and includes a taxonomy which represents geographical entities with physical-natural aspects. All the classes of GO-PHY are sub-classed of *astronomical entity*, *physical entity*, *geographic entities*, *natural entities*, *event* and *terrestrial entity* classes of GO-TOP.

GO-HUM contains 204 classes, 8 object properties. It imports the GO-TOP module, and is organized in a taxonomy which constitutes an inventory of geographical entities created by humans. The high level classes imported from GO-TOP are *astronomical entity*, *anthropic entity*, *geographic entity*, *event*, *go entity*, *length*, *non-physical entity*, *physical entity* and *terrestrial entity*, from which GO-HUM defines its specific subclasses. The main specific object properties are: *fought between*, *composed by*, *has stop over*, *has length*, *has path*, *has cultural heritage of* and *won*.

GO-FAR contains 87 classes, 2 object properties. It imports the GO-TOP module, and describes all (and only) the geographic features (including places, people and events) produced by human during ancient times, with particular reference to ancient Rome as the main scope of this ontology is the annotation of Latin texts. Moreover, it includes, among others, some specific entities and classes which describe the Ancient World imported from *ancient entity*, *socio-institutional entity*, *group of people*, *populated place*, and *artifact* classes of GO-HUM, *geographic entity* from GO-TOP. Finally, it has *has real place* among the Object Properties.

An example of the connection of these modules might be represented by the class of *GO! Entities*, which includes, among others, *Natural Entities*, *Anthropic Entities*, *Astronomical Entities*, *Physical Entities*, *Terrestrial Entities* and *Unreal Entities*, as can be seen in Fig. 1.

Each of these subclasses of *GO! Entities* contains further subclasses and is characterized by specific properties and relations (*name*, *location*, *length*, *size*, *spatial relation* and so forth). In this sense, the *GO!* modelling choices allow to express a range of information about geographical places (i.e. their evolution through time as attested

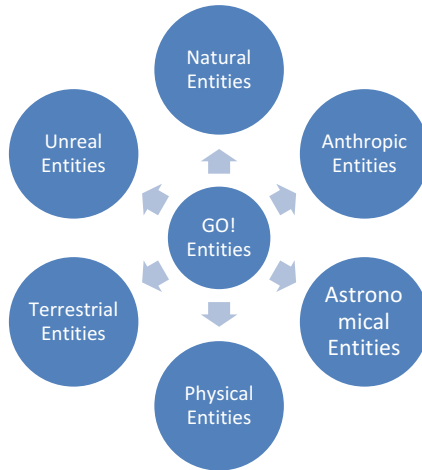


Fig. 1. GO! entities

by the texts which mention them, GPS coordinates, physical and geopolitical descriptions, switch of name, and so forth); they also allow to describe historical events connected with specific places. Connections with places data available in Pleiades are used; moreover GO! can manage imaginary places; the Open Annotation ontology is adopted to cite the passages.

5 Envisioned Uses

In the digital humanities field a geographical turn happened, whose meaning is that the geography is seen no more simply as a specific discipline rather its meaning is that of the environment where most of the activities and knowledge lie. A glue, a substrate connecting most of what exists and happens in the human world [4]. The Geolat project is a product of this idea.

In fact, in Geolat the annotation of texts based on the ontology is conceived for at least two different scopes:

- browsing and searching texts on the basis of the internal geographical content and knowledge, made visible and usable thanks to the ontology;
- production of digital (online) and printed critical editions of new type: geographical critical editions.

The first scope can be exemplified saying that the reader draws an area on a map and the system shows on the map the places mentioned in the texts of the digital library underlying the map. That is the interconnection between the library of digitally annotated texts and the map allows for a truly exploratory (the reader has not to know anything very precise about the area s/he are investigating), yet controlled (the reader can fine tune the texts collection which is analysed: period, authors, genres, ...) approach to a given collection of texts. But the same cartographical query can produce lists of authors, works, passages, mentioning places belonging to the area traced by the reader on the map. Or inversely the reader starts from one or more works and can display on the map the places mentioned in those works. The apparently simple scope of searching can become very powerful if one considers that in the annotation of places events and involved people with their roles can be described. This way one can search e.g. for the geographical places where *proconsules* are mentioned. This is not possible with more simple NER approaches where the place name is a string and not - as it happens instead when using an underlying rich ontology - a complex entity of meaning full of connections and implications.

In the second scope the concept of “digital edition of a text” is developed/expanded. The “edition of a text” has today - at least for classical works, but many things are similar for the editions of medieval or modern texts - two main forms: with or without the description of the variants and their evaluation by the curator of the edition in the *apparatus criticus*. In the first case the edition is usually qualified as a *scholarly edition*, in the second one it is usually called a *variorum edition*. In both cases the editor of the texts offers what he judges being the best reconstruction of the true text issued by the hand of the author. And this type of work on the text was consolidated through thousands of years (the first critical editions of texts were produced more than 2000 years ago). The scope of a scholarly edition (be it of the first or of the second type just described) is that of offering those who want to study a given work a rich informational environment. Until these years the main type of information offered is that of an accurate description of the state of the manuscript tradition. Nowadays, tools are available allowing to expand the type of critically assessed information which is offered to the researchers who study a given text, what is both the cause and the effect of new study perspectives. For example, we are *more aware* of the fact that the geography conveys political and ideological meanings: giving names to specific points on the surface of Earth means affirming a property on them; the same is for defining boundaries; and mentioning specific place names in groups is a way of declaring interpretations of facts, or of affirming political or religious positions. This type of research interest needs the support of an edition of the text giving space to this study approach: a geographer will probably appreciate a scholarly edition enriched by a geographical apparatus that is an edition with a good comment of the text plus a full spectrum of documentation about the geographical knowledge contained in the text that is link the source describing in all the possible ways the places which are mentioned (Fig. 2):

- 1 Se quoque, cum transiret mare, non Ciliciam aut Lydiam—quippe tanti belli exiguam hanc esse mercedem—sed Persepolim, caput regni eius, Bactra deinde et Alexandria ultimique Orientis oram imperio destinasse. Quocumque ille fugere potuisset, ipsum sequi posse: desineret terrere fluminibus, quem sciret maria transisse.
- 5 Reges quidem haec invicem scripserant. Sed Rhodii urbem suam portusque dedebant Alexandro. Ille Ciliciam Socrati tradiderat, Philota regioni circa Tyrum iusso praesidere. Syriam, quae Coele appellatur, Andromacho Parmenio tradiderat bello, quod supererat, interfuturus.

-
- 1 Ciliciam: [<http://www.geonames.org/8378491/cilicia.html>]
Lydiam: [<http://www.trismegistos.org/place/1269>]
- 2 Persepolim: [<https://pleiades.stoa.org/places/922695> | <https://vici.org/vici/20445/> | <http://gazetteer.dainst.org/place/2043479>]
Bactra: [<https://pleiades.stoa.org/places/961886>] also known as Zaraspadum and Zariaspa
Alexandria: Alexandria MSA [Alexandria of Egypt <http://pleiades.stoa.org/places/72707> | Alexandria Eschate <http://pleiades.stoa.org/places/59672>], Alalia MSB [<https://pleiades.stoa.org/places/472048>]
- 5 Rhodii: inhabitants of Rhodos [<https://pleiades.stoa.org/places/590030>]
- 6 Ciliciam: see 1.1
Tyrum: Tyrum MSB [<http://pleiades.stoa.org/places/609564>], Tyraion MSC [<https://pleiades.stoa.org/places/609564> | <http://dare.ht.lu.se/places/21528>]
Syriam ... Coele: [<https://pleiades.stoa.org/places/991407> | <http://www.geonames.org/8378530/syria-coele.html>]

Fig. 2. Mock-up of a scholarly geographical edition

while a classical scholar will probably appreciate a scholarly variorum edition enriched by a geographical apparatus, that is an edition where it is possible to cross-reference the variants of the manuscript tradition with the geographical knowledge. But other types of information could be equally interesting: historical information about the events mentioned in the text, or prosopographical information about the persons involved in the events, or both types merged. No single printed edition, nor single scholar could bear the weight of this complexity:

(variorum: yes/no) AND/OR (geographical: yes/no) AND/OR (historical: yes/no) AND/OR (prosopographical: yes/no)

which can probably only be managed if the edition is conceived as a collaborative work of different scholars and gives origin to an edition whose content typology is dynamically generated on the basis of the researcher's interest (strictly philological AND/OR geographical, etc.). In other words we suggest here that the complexity of the digital edition be conceived not at the very philological level only (as variorum edition, with complex discussions about what it must contain and how the content must be presented also in rapport with the printed version of this type of edition) but also as a research tool which creates a research environment which configures itself according to the scope and the interests of the researcher/reader (Fig. 3).

All the above is possible because the specific value of a digital edition is not only in the digital form of representation of textual information: dynamic rather than static, resulting in better visual or practical usability; but it mainly lays in the ability to work with computational methods on the text and on the information it conveys [12, 13].

In conclusion, the digital edition of a text should aim to provide adequate data and functionality to further forms of processing. Hence the idea that the “digital scholarly edition” until now often identified with the “digital critical edition” also known as “digital variorum edition”, can also take other forms focused on other types of ‘scholarly research’: from the geographical knowledge contained in the text, to the

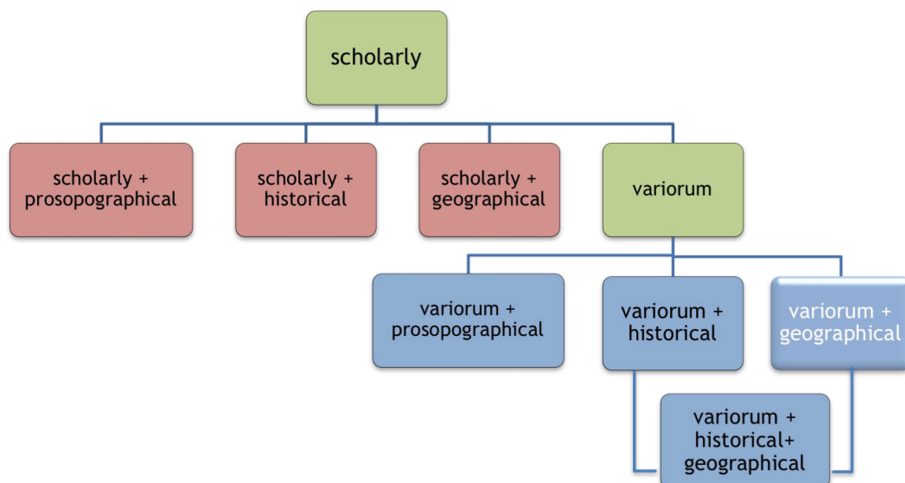


Fig. 3. Various types of edition of a text: actual (green) and foreseeable (other colors) (Color figure online)

historical knowledge (time and events) often inextricably linked with the prosopography, and much more.

So, if the digital critical edition (digital variorum edition) is a type of digital scholarly edition containing an apparatus that analyses and describes the state of the text in the witnesses, then we can conceive e.g.

- the digital scholarly geographical edition of a work – whose apparatus contains an analytical description of the geographical knowledge contained in the place names;
- the digital critical geographical edition (digital variorum geographical edition) whose geographical apparatus is layered over a base critical edition:

To do so the knowledge contained in the text must be expressed in a highly formal manner - the same way that the critical apparatus is a highly formal device – and an ontology is a good and *very complex* means to do that. Here below an abstract sample of a passage of text where the place name Lydia is annotated with reference to the GO! ontology (Fig. 4):

1 Se quoque, cum transiret mare, non Ciliciam aut <geogName ref=https://w3id.org/geolit/ontologies/GO-PHY 'Lydia' xml:id='Lydia'>Lydiam</geogName> – quippe tanti belli exiguum hanc esse mercedem—sed Persepolim, caput regni eius, Bactra deinde et Alexandria ultimique Orientis oram imperio destinasse. Quocumque ille fugere potuisset, ipsum sequi posse: desineret terrere fluminibus, quem sciret maria transisse.

Fig. 4. Passage where the word Lydiam is annotated

More references to our studies in this field can be found also in forthcoming publications, namely the proceedings of the conferences DHANT (MSH-Alpes, Grenoble 2015) and Dixit (Köln 2016).

References

1. Bianchetti, S.: *Geografia storica del mondo antico*. Monduzzi, Bologna (2008)
2. Casati, R., Smith, B., Varzi, A.C.: Ontological tools for geographic representation. In: Guarino, N. (ed.) *Formal Ontology in Information Systems*, pp. 77–85. IOS Press, Amsterdam (1998)
3. Casati, R., Varzi, A.C.: *Parts and Places*. MIT Press, Cambridge (1999)
4. Elliott, T., Gillies, S.: Digital geography and classics. In: *DHQ: Digital Humanities Quarterly*, Changing the Center of Gravity: Transforming Classical Studies Through Cyberinfrastructure, vol. 3(1) (2009). <http://www.digitalhumanities.org/dhq/vol/3/1/000031/000031.html>
5. Frank, A.: Spatial ontology. In: Stock, O. (ed.) *Spatial and Temporal Reasoning*, pp. 135–153. Kluwer Academic, Dordrecht (1997)
6. Geus, K., Thiering, M. (eds.): *Feature of Common Sense Geography. Implicit Knowledge Structures in Ancient Geographical Text*. Lit Verlag, Wien-Berlin (2014)
7. Guarino N.: Formal ontology and information systems. In: *Proceedings of FOIS 1998*. Trento, Italy, 06–08 June 1998, pp. 3–15. IOS Press, Amsterdam (1998)
8. Kavouras, M., Kokla, M., Tomai, E.: Comparing categories among geographic ontologies. *Comput. Geosci.* **31**(2), 145–154 (2005)
9. Kuhn, W.: Ontologies in support of activities in geographical space. *Int. J. Geogr. Inf. Sci.* **15**(7), 613–631 (2001)
10. Lana, M., Borgna, A., Ciotti, F., Tambassi, T.: Ontologies and the cultural heritage. The case of GO! In: *Proceedings of the Second International Workshop on Semantic Web for Scientific Heritage Co-Located with 13th Extended Semantic Web Conference (ESWC 2016)* Heraklion, Greece, 30th May 2016, *CEUR Workshop Proceedings*, vol. 1595, pp. 7–18 (2016). urn:nbn:de:0074-1595-0, ISSN 1613-0073
11. Momigliano, A.: *Alien Wisdom. The Limits of Hellenization*. Cambridge University Press, Cambridge (1975)
12. Monella, P.: Why are there no comprehensively digital scholarly editions of classical texts? In: *IV Meeting of Digital Philology*, Verona, 15 September 2012. http://www1.unipa.it/paolo.monella/lincei/files/why/why_paper.pdf
13. Pierazzo, E.: *Digital Scholarly Editing. Theories, Models and Methods*. Ashgate, Farnham Burlington (2015)
14. Smith, B.: Ontology. In: Floridi, L. (ed.) *The Blackwell Guide to the Philosophy of Computing and Information*, pp. 155–166. Blackwell, Malden (2004)
15. Smith, B., Mark, D.M.: Ontology and geographic kinds. In: Poiker, T.K., Chrisman, N. (eds.) *Proceedings of the Eighth International Symposium on Spatial Data Handling*, pp. 308–320. International Geographical Union, Burnaby (1998)
16. Smith, B., Mark, D.M.: Geographical categories: an ontological investigation. *Int. J. Geogr. Inf. Sci.* **15**(7), 591–612 (2001)
17. Tambassi, T.: Rethinking geo-ontologies from a philosophical point of view. *J. Res. Didactics Geogr. (J-Reading)* **2**(5), 51–62 (2016). <https://doi.org/10.4458/7800-04>
18. Tambassi, T., Magro, D.: *Ontologie informatiche della geografia. Una sistematizzazione del dibattito contemporaneo*. *Rivista di estetica* **58**, 191–205 (2015)
19. Turner, A.J.: *Introduction to NeoGeography*. O'Reilly Media Inc., Sebastopol (2006)

A Research Tool for the ERC-Funded EMOBookTrade Project

Giliola Barbero^{1(✉)} and Luigi Tessarolo²

¹ Università degli Studi, Udine, Italy
giliola.barbero@uniud.it

² Venice, Italy

Abstract. The ERC-funded EMOBookTrade project, led by professor Angela Nuovo and based at the University of Udine, addresses the issues of book prices and of book privileges in early modern Europe (1540–1630c.). To achieve the goals set, a Web application is under development by the research team. This paper describes technical solutions found to ensure usability and effectiveness in recording complex historical data, and automatisms which allow data analysis.

Keywords: Digital humanities · Digital history · History of the book
Price history · Economic history

1 The ERC-Funded EMOBookTrade Project

The Early Modern Book Trade project, funded by the European Research Council, aims to understand how the European book market worked during the sixteenth century and the first decades of the seventeenth [1]. The project team is analyzing prices of books, privileges granted by Western governments and any other evidence concerning bookshops management and commercial networks.

Until this moment the history of the book has taken into account cultural and technical issues related to the production and circulation of books, while economic and juridical aspects received little attention. Nevertheless, several innovative researches show how important the booksellers' economic organization was in spreading texts and ideas, not only in Europe but also in the new world [2, 3].

To understand how economic and juridical issues influenced the cultural weight of printed books, the EMOBookTrade project is developing five main activities. (1) One research activity is devoted to investigate prices set by Italian publishers and booksellers (Manuzio, Giolito, Scoto, Compagnia bresciana) and to study the inventory of Gian Vincenzo Pinelli's private library, as it was assessed for sale during an auction that took place in Naples in 1608. (2) A second research activity on European prices employs Christophe Plantin's archive in Antwerpen as well as French publishers' and booksellers' trade lists. (3) A third task focuses on the management of a Venetian bookshop

L. Tessarolo—Independent Scholar.

and the pricing of books in the early seventeenth century, and consists in the transcription and analysis of Bernardino Giunti's stock book. (4) The census, edition and interpretation of privileges granted by the Republic of Venice and preserved in the Archivio di Stato in Venice are underway. (5) The correspondence between Venetian wholesale bookseller Giovanni Bartolomeo Gabiano and his European business partners, which describes the technique of building and managing a transnational network of book distribution is in the process of being edited.

To manage these sources and related data, a database application has been developed, in collaboration with the project team and Luigi Tassarolo, an IT specialist with a long standing experience in the field of digital humanities [4].

The database application consists of a backend and a frontend. Through the backend (which has already been working for six months) the EMOBookTrade research team enters data and manage controlled vocabularies and authority files. Researchers can also formulate complex queries to help data interpretation. The frontend is still under development: it is intended to publish data and make them accessible to the scholars outside the project. Both the backend and the frontend are in English.

The reliability of prices analysis depends not only on the accuracy of research activities and on a correct historical interpretation of the sources, but also on the semantic evaluation of data and of their relations inside the database (DB) conceptual model. In a web application for an innovative historical research, the semantic value of each entity must be defined precisely, because there are not always existing standards to adopt.

But implementing a common language and a correct data classification is not the only task when designing a DB application for humanistic researchers. Also usability and effectiveness of the Database Management System (DBMS) are fundamental, because researchers do not want to spend time in doing what they do not need to do while working on traditional paper notebooks or on word processors.

Finally, the EMOBookTrade project and the related DB application far are proving that the effectiveness of a digital repository of historical sources does not rely only on digital images but also on their transcription and interpretation, a task that only a research project can afford.

This paper aims to describe the DB (Sect. 2) and highlight the problems—not only technical—encountered when designing a Database Management System (DBMS) for historical research (Sect. 3); while—as the frontend is still in progress—only few information will be given on it (Sect. 4).

2 Entities in the DB

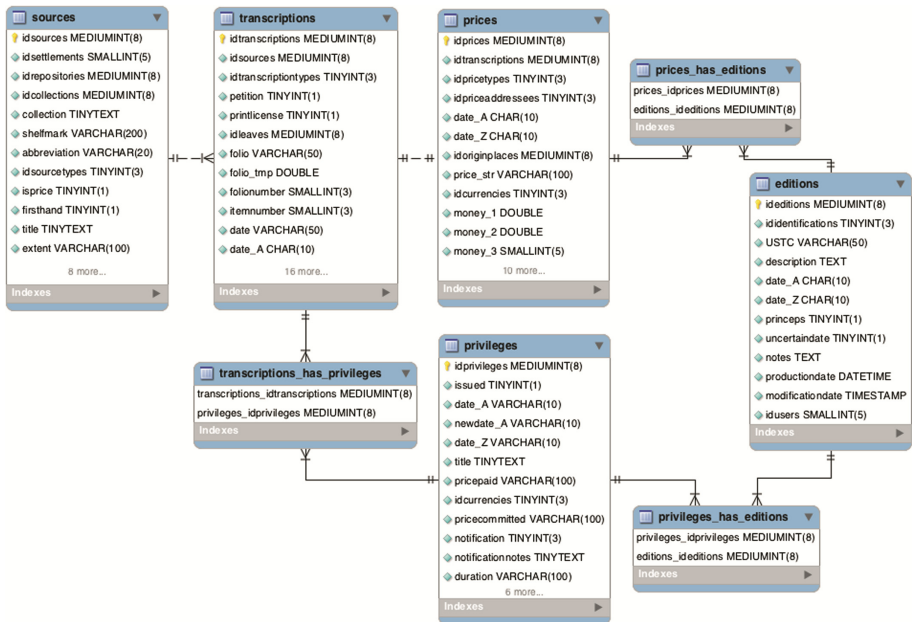
The EMOBookTrade DB and its corresponding DBMS aim to publish sources and to support their interpretation and analysis. The team needs to discover and enquiry whether there were constants in pricing books, what the average prices were, whether and how prices change over time and which elements could prompt publishers and booksellers to set different prices from the average (lower or higher).

To evaluate these issues, first of all a great effort was made to design a valid conceptual model.

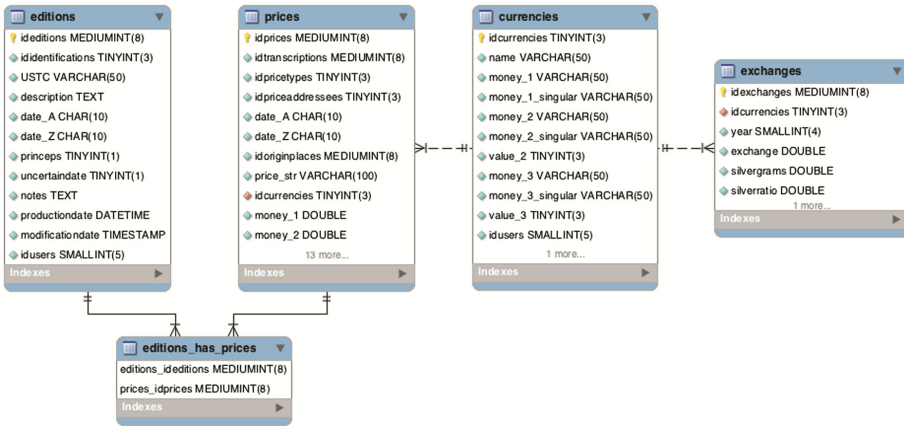
In the EMOBookTrade DB the main entities are the following:

- sources: publisher catalogues, bookshop and bookstore inventories, archive documents;
- transcriptions: texts transcribed from the sources that provide information on prices and privileges;
- prices: an amount of money indicated in one or more sources;
- privileges: a public act described in one or more sources;
- editions: a series of printings of the same book priced and/or privileged by one or more sources;
- names: personal, corporate or geographic names, e.g. authors, publishers, places of publication, booksellers;
- currencies in prices, e.g. Lire veneziane, Ducati napoletani, Livres tournois.

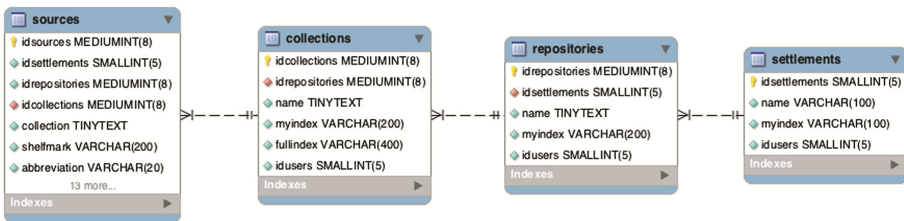
Items of this list represent the historical objects and events considered by the EMOBookTrade project at an abstract level, while the logical structure of the tables is more complicated. Here is a simplified illustration of the fundamental tables and their relationships:



Other DB tables contain price-specific information which is one of the two cores (the other one is about privileges) of the EMOBookTrade project: prices, currencies and exchange rates are linked to editions as follows:



The attributes of each entity have been designed to respect standard, at least when existing. Sources identifier data (source location) for example respect the TEI standard for manuscript description (Settlement, Repository, Collection, Shelfmark) [5]:



And even if data about editions do not respect Marc [6] and Unimarc [7] models (because they would be too complicate for a non-bibliographic DB), each edition is identified by a public ID derived from institutional repositories such as the *Incunabula Short Title Catalogue* (ISTC) [8] and the *Censimento delle edizioni italiane del XVI secolo* (EDIT16) [9].

To design the conceptual model (entities) and the data structure (tables and relationships), as a first step, the team discussed and listed ‘what’ they wanted to record and, in some cases, what they had already listed in previous DBs. In fact a DB of Plantin’s archive and a DB of privileges were already in use by some members of the new team. From these previous DBs many fundamental ideas came to the EMOBookTrade project, but afterwards a great effort has been made to restore those DBs structures and to design new tables and their relationships.

DBs developed for individual research in the field of humanities in fact often carry some recurring mistakes, that are tolerable in a private context but cannot be managed in a shared public digital repository.

One of the most common mistakes is to consider as one single entity’s attributes sets of data that do not have a one-to-one ratio. E.g. librarians and IT specialists working for libraries with bibliographic formats (Marc, Unimarc) know very well that the relationship between authors’ names and bibliographic descriptions is a many-to-many one, but

in several individual research contexts bibliographic descriptions and authors' names are stored in two fields of the same table and when authors are more than one, different names are copied together in the same field. The consequence is that each record can be read by human eyes as on a page on paper, but data as a whole cannot be ordered and managed automatically.

A too much scant attention to the relational structure can be observed also in other common practices. For example few pre-set tables with controlled vocabularies are designed and free fields are preferred when any sort of classification (classification of source, editions...) has to be recorded, that does not help in retrieval.

Moreover another common mistake in individual DBs is to design a unique field to store ontologically different data. For example it can be a discouraging experience to explain to humanistic researchers that the 'responsibility' a person has in a specific context (e.g. the responsibility to be the copyist of a document) is not to be considered a fixed 'qualification' of that person, because the same person can be a copyist in a context and an author in another one.

Last but not least also recording different kind of information in a field designed for a specific typology of data, just because 'there is a blank to fill up', is a typical misunderstanding of individual research tools. This is again a sort of typical confusion between a DB record and a traditional page, where the responsibility of understanding the semantic value of data and their internal links falls on the human reader.

On the other hand, from the previous DBs already in use, the new EMOBookTrade DB derived a great set of specialist data collected during the previous research, which saved the team from the task of speculating an abstract model and the entities which could be proved useful in the future. As the design of a DB devoted to an historical research project cannot follow existing standardization - for its innovative nature -, previous experience and data have been fundamental to design the EMOBookTrade DB.

3 Usability and Effectiveness in the DBMS

After designing the logical model and creating the data structure, in developing the DBMS, the main task was to put the team in the condition to operate efficiently through the backend. The backend was organized in the safest and most expedient possible way, even if these two needs can often conflict. To this regard, balanced solutions were found, and at present the EMOBookTrade team members are working in an suitable condition.

The DBMS complies with the principles outlined below.

- No data has to be saved through the existing 'save' buttons, but when an empty record is closed, a pop-up message appears asking to the client if that record should survive or not.
- No data should ever be entered twice. E.g. when a researcher is transcribing a series of prices from the same inventory, and each price is intended to populate a single record, a basic set of data shared by all the prices—e.g. the date, the typology of price, the place where that price was set - is inherited (although they can also be changed manually).

- When information is already available online in free access repositories, such information can be imported automatically minimizing manual interventions. For this reason old books are not being catalogued by the EMoBookTrade team. Instead bibliographic descriptions are imported automatically from existing national and international catalogues and bibliographies such as USTC [8], EDIT16 [9], and VD16 [10].
- Numeric series (such as the order number of pages of a single catalogue) are entered automatically, although a manual change option is also available.
- Researchers are not required to perform manual calculations, because multiple automated functions are made available instead. For example, for each edition considered, the EMoBookTrade DBMS calculates the price per printing sheet. This is made possible by a function that automatically divides the total price found in the source by the number of leaves, divided in turn by the volume's format. For example, if a given book carries the total price of 1560 denari and it is known to be an in 4° of 360 leaves (for a total of 90 printing sheets), the price per printing sheet will be 17.33 *denari*. Moreover each price is automatically calculated in the minimum sub-multiple of the original currency used by the source, then converted in Venetian *denari* (when different) and in the correspondent grams of silver:

RECALCULATED PRICE

Total price	Deniers	20	Venetian denari	57.1587	Grams of silver	1.3828
Price per sheet (5.5)	Deniers	3.6364	Venetian denari	10.3925	Grams of silver	0.2514

Calculate the price per sheet based on the information derived from the linked edition

The conversion from foreign currencies to Venetian *denari* and from *denari* to grams of silver also takes into account the date when pricing was set, because the chronological fluctuation in exchange rates are also recorded in the DB.

- Any data resulting from complex serial calculations is a dynamic entity and is not to be permanently fixed and stored within the DB. Any manual intervention made on a single numeric value will result in a consequent automatic readjustment of all the interlinked values. E.g. if a recorded total price needs to be corrected, because at a first glance the researcher misread the original source, the correspondent price per printing sheet will be automatically recalculated, and the same goes for exchange rates and so forth.

Moreover several data, as they result from the interpretation of historical sources, cannot always be taken as fully reliable. For example sources may be damaged, misleading or carry incoherent information.

At times, data can even be ‘unidentified.’ This, far from being a paradox, is a commonplace of historical bibliography. This is the case when a source cites a work without providing sufficient bibliographic data: e.g. ‘Rime di Petrarca’. The given case does not allow to establish an univocal relation between the cited - and priced work - and a known edition. In order to store information that are still valued by the researcher, the information regarding the price of the cited work will be linked to a fictitious bibliographic record (unidentified edition) based on the scanty information provided by the source: e.g. ‘Francesco Petrarca, Rime’.

Appropriate models of data, usability and effectiveness of the DBMS and the freedom to record also doubtful data, help researchers engaged in digital humanities to accept the introduction of new shared digital tools in their traditional environment. Without care of such issues, it is very difficult to develop and employ DBs and DBMSs in the field of humanities and in humanistic projects [11].

4 The Future: Data Analysis

While the DB and the backend system described here are already working, the frontend - the public web site - is being designed. Nevertheless, in the backend the possibility to manage data for individual research already exists. At present, prices can be ordered and sorted by total price, price per printing sheets or quantity of printing sheets, and, most importantly, editions can be searched to verify if they had received privileges and to find out what prices were set by publishers or paid by customers for them. Data can also be downloaded by the team researchers.

This module of the backend has already been exploited to elaborate data for the congress *Selling & Collecting. Printed Book Sale Catalogues and Private Libraries in Early Modern Europe* held in Cagliari in September 2017 [12].

In the meantime the frontend is under construction. The EMOBookTrade team is discussing different solutions in order to create a research engine suitable to the scholars and their scientific needs. The starting point of the main public search will be from editions, to obtain prices and privileges connected with them. Scholars will be able to answer questions like “How much did Petrarca’s *Canzoniere* cost in Venice during the 15th century?”, or “Which edition of Ariosto’s *Orlando furioso* received a privilege in Italy?”. Moreover they will be able to obtain and compare prices in any currency and/or in grams of silver and to order prices and privileges by different criteria.

Due to the analytical conceptual model designed, and to the amount of data collected by researchers, we trust in a coherent and suitable future development of the frontend. Just because the EMOBookTrade researchers have striven to share a common language and a common data classification, scholars outside the project will be able to share the same logic and consequently to discuss and exploit scientific results of the project.

References

1. EMOBookTrade, An Evidence-based Reconstruction of the Economic and Juridical Framework of the European Book Market, <http://emobooktrade.uniud.it/>
2. Nuovo, A., Ammannati, F.: Investigating book prices in early modern Europe: questions and sources. *JLIS.it* **8**, 3, September 2017. <https://doi.org/10.4403/jlis.it-12365>

3. Musisque Deoque, Un archivio digitale di poesia latina, PRIN 2005 e 2007, progetto scientifico di P. Mastandrea, R. Perrelli, G. Biondi, L. Zurli, V. Viparelli, progetto informatico di L. Tassarolo. www.mqdq.it. *Pede certo - Metrica Latina Digitale*, Università di Udine, progetto scientifico di L. Mondin, progetto informatico di L. Tassarolo, 2013. www.pedecerto.eu. Pietro Metastasio, *Drammi per musica*, PRIN 1998, progetto di A.L. Bellina e L. Tassarolo. www.progettometastasio.it. Carlo Goldoni, *Drammi per musica* PRIN 2001, progetto di A.L. Bellina e L. Tassarolo. www.carlogoldoni.it. *Varianti all'opera*, FIRB 2008–2012, Università di Padova, di Milano e di Siena. www.variantiallopera.it. Miguel de Cervantes, *Novelas ejemplares*, PRIN 2004, progetto scientifico di D. Pini; progetto informatico di L. Tassarolo. www.cervantes.cab.unipd.it. AXON - *Iscrizioni storiche greche*, Università Ca' Foscari Venezia, coordinatrice S. De Vido, supervisore C. Antonetti, progettazione e realizzazione S. Palazzo, M. Socal, L. Tassarolo (progetto informatico) (2015). <http://virgo.unive.it/venicepigraphy/axon>
4. Squassina, E.: *Authors and the System of Publishers' Privileges in Venice*. *Gutenberg Jahrbuch*, 42–74 (2016)
5. TEI – Consortium. Text Encoding Initiative, <http://www.tei-c.org/index.xml>
6. MARC standards. Library of Congress: Network Development and MStC standards Office, <https://www.loc.gov/marc/>
7. UNIMARC Bibliographic, <https://www.ifla.org/publications/unimarc-bibliographic-3rd-edition-updates-2012?og=33>
8. USTC Universal Short Title Catalogue, <http://www.ustc.ac.uk/>
9. EDIT16 Censimento nazionale delle edizioni italiane del XVI secolo, http://edit16.iccu.sbn.it/web_iccu/ihome.htm
10. VD 16 Verzeichnis der im deutschen Sprachbereich erschienenen Drucke des 16. Jahrhunderts, https://opacplus.bib-bvb.de/TouchPoint_touchpoint/start.do?SearchProfile=Altbestand&SearchType=2
11. Vitali, S.: *On DB and historical research*, *Passato Digitale*. Bruno Mondadori, Milano (2004)
12. *Selling & Collecting. Printed Book Sale Catalogues and Private Libraries in Early Modern Europe*, <http://convegni.unica.it/icsc/>

The Biographical Dictionary of Friulians - “Nuovo Liruti” Online

A Biographical Dictionary Based on Semantic Web and Linked Open Data

Stefano Allegrezza¹(✉) and Nicola R. Di Matteo²(✉)

¹ Università degli Studi di Udine, Udine, Italy
stefano.allegrezza@uniud.it

² Dalhousie University, Halifax, NS, Canada
nicola.dimatteo@dal.ca

Abstract. Most biographical dictionaries today are becoming more and more online based. For example, Treccani has produced a digital edition of the Italian Biographical Dictionary, which is more readily available and more up-to-date than the paper edition. In fact, the traditional edition, which is organized according to an alphabetical order, has the entries relating to the letter “A” published in 1960 and those related to the letter “M” published in 2011; therefore it is currently obsolete. The Rosi Dictionary of the Renaissance Risorgimento is fully accessible online, as well as the Biography of Women and Men of Resistance of the National Association of Partisans of Italy, the Online Biography of Protestants in Italy of the Society of Studies Valdesi and many other examples that could be made. A similar trend amongst these is the American National Biography, the Diccionario Biográfico Español, the Slovenska Biografija show, only to make some significant examples. However, these online dictionaries often don’t have an intuitive interface and often are not so “attractive” to entice the user into navigation. Moreover, semantic web tools are almost never used, but they can find their own special and interesting application in this particular field. The aim of this project was to realize not only a digital edition of the printed version of the “Biographical Dictionary of Friulians” (“New Liruti online”), but also one of the richest and structured cultural and historical information deposit on Italian web sites, thanks to the application of semantic web methodologies for the digital edition of biographical dictionaries, with the opportunity to reach a much more ambitious and potentially unlimited audience than the paper edition - as well as being one of the most important cultural initiatives within the broadest project on “Cultural Identity of Friuli (ICF)”. It is also an example of effective collaboration between humanists and computer scientists, without whom the project would not be possible.

Keywords: Biographical dictionaries · Biographies · Semantic web
RDF · Linked data

1 Introduction

Most biographical dictionaries now tend to become more and more online. For example, Treccani has produced a digital edition of *the Italian Biographical Dictionary* [1], which is more readily available and more up-to-date than the paper edition (which is organized according to an alphabetical order, with the entries relating to the letter “A” published in 1960 and those related to the letter “M” published in 2011), is currently obsolete; the *Biography of Women and Men of Resistance* [2] of the National Association of Partisans of Italy, the *Online Biography of Protestants in Italy* [3] of the Society of Studies Valdesi and many others examples that could be made. Abroad the trend is almost identical, as *the American National Biography* [4], the *Diccionario Biográfico Español* [5], the *Slovenska Biografija* [6] show, only to make some significant examples. However, these online dictionaries have often a not intuitive interface and often not so “attractive” to entice the user into navigation. Moreover, semantic web tools are almost never used, but they can find their own special and interesting application in this particular field. This project aims to be a best practice for the application of innovative methodologies in the realization of digital editions of biographical dictionaries, and also an example of collaboration between humanists and computer scientists, without which the project would not have been possible.

2 The Idea of the Project

The idea of the project started with the happy intuition of the Pio Paschini Institute for the Church’s history in Friuli, which in 2016 proposed to create a digital edition of the “New Liruti. Biographical Dictionary of the Friulians”¹ [7] in collaboration with the institutions that had promoted the printed edition. The entries already published would have been revised and integrated by about four hundred bio-bibliographic profiles of the so called “Ongaro Supplement” (made by Maiko Favaro on the basis of the eighteenth century manuscripts by Domenico Ongaro), and by the voices of Onomasticon, which was planned during the presentation of the whole work. It was therefor not a question of simply transposing the digital version of the printed edition of the “New Liruti” or of making its electronic version available (the PDF file) but something much more: a real online biographical dictionary with a captivating graphic interface with numerous functionalities that can be used both by the scholar and the simple citizen who want to deepen the history and culture of Friuli Venezia Giulia. The project was funded by the Province of Udine, the Friuli Foundation and the Archdiocese of Udine; it was promoted by the Pio Paschini Institute for the Church History in Friuli, the Patriarch Deputies for Friuli, the Historical Institute of the Ancient Book, the Friulian Philological Society and the Department

¹ Although spellings Friulian and Friulan are both used, Friulian is more used and acknowledged [9, 10].

of Humanities and Cultural Heritage of the University of Udine². The scientific directors are Cesare Scalon and Claudio Griggio while the technical director is Stefano Allegrezza. The work was entrusted to Nicola Raffaele Di Matteo. The project was officially presented on April 3, 2017, Friuli’s home day, and is now available at the address <http://www.dizionariobiograficodefriulani.it> (see Fig. 1).



Fig. 1. The home page of the Biographical Dictionary of Friulians

3 Strengths of the Project

The project has started from a preliminary phase of analysis of biographical dictionaries published on the web, both in Italy (such as the Italian Dictionary of Biography, the Rosi Dictionary of the Risorgimento Renaissance, the Biography of Women and Men of Resistance by the National Association of Partisans of Italy, the Online Biography of Protestants in Italy by the Society of Studies Valdesi, etc.) and abroad (such as the American National Biography, the Deutsche Biographie, the Slovenska Biografija and the Diccionario Biográfico

² Among the partners of the project are the Accademia San Marco in Pordenone, the Academy of Letters and Arts in Udine, the Diocesan Archives and Patriarchal Library of Udine, the State Archives of Gorizia, the State Archives of Pordenone, the State Archives of Udine, the Guarneri Civic Library of San Daniele del Friuli, the Civic Library of Pordenone, the Civic Library “Vincenzo Joppi” in Udine, the Isontina State Biological Bureau of Gorizia, the “Institute of Social and Religious History” of Gorizia, the popular University of Udine.

Español, only to make some significant examples) [15]. From this analysis we have become aware of the almost universal trend of printed biographical dictionaries to become online biographical dictionaries. The reasons behind this tendency are many and will be briefly examined in the following because they are the same that are at the basis of this project.

3.1 Continuous and Real-Time Update

Compared to the printed edition, the online edition has the undoubted advantage of providing an up-to-date and continuously updated reference. Once a new biographical entry is inserted, it will be immediately visible online. Even correcting any mistakes or denials is very easy and immediate (this is obviously not possible with the printed edition). Anyway this requires the presence of an editorial committee that continuously follows the editorial activities.

3.2 Better Usability

Usability of content that is made available online could be better than the printed edition, since it is possible to not only a sequential reading mode (such as the printed edition) but also a hypertextual reading mode (taking advantage of the links that were included in the text to highlight the most interesting links). In addition, the 2620 biographical entries, re-checked and updated where necessary, are available in multiple navigation modes: not only following an alphabetical order - as is evident - but also following chronological, geographic or thematic paths.

3.3 Unlimited Scalability

The online edition of the dictionary is based on an infrastructure that puts no limit to the variety and amount of information that can be hosted; the dictionary can be expanded with content of any kind (think, for example, about audio recordings that can be associated with a musician's biographical entry or video recordings that can be associated with a director's entry, etc.). There are basically no limits to the expansion of the online dictionary.

3.4 Possibilities of Interaction with Users

The electronic medium allows extremely varied forms of interaction with users and in some cases also suitable for content editing, as it is possible, for example, in Wikipedia. Although mechanisms of this kind are widely used with good results [8], it could be necessary to have a drafting committee to verify the content inserted by users.

As a result, it has been decided to allow users to interact with the site only by leaving comments on biographical entries. It is also possible to interact with major social media (Facebook, Twitter, Google+, Instagram, LinkedIn). In the future it will be possible to review this choice by enabling users to interact with the site by

adding materials that can enrich a given biographical entry. For example, think of an artist’s biographical entry and the likelihood that a user who has digital copies of some of his works can add them, thus enriching the biographical entry with potentially interesting material. This will require the definition of “strategies” to check and “evaluate” the content by an editorial committee.

3.5 Possibility to Carry Out Very Sophisticated Searches

The true richness of the digital edition is certainly the ability to carry out searches among the most disparate and sophisticated. Compared to the printed version, which basically allows you to search only alphabetically by browsing the 7285 pages of printed volumes one by one.³ The digital version allows you to easily and freely perform both full-text searches on all content [11], that advanced searches (by specifying the appropriate search criteria so that you can quickly get the content you want to see). To achieve this, all biographical entries have been associated with a series of metadata (place and date of birth, place and date of death, places and dates important in person’s life, profession, sex, curator of biographical voice, etc.). This allows the user to retry information about persons or the facts through different search criteria that have been defined; for example query by name and surname, date or place of birth and death, profession (jurists, literate, typographers, musicians, etc.), sex, etc. Therefore, it is possible to know which illustrious Friulians are related to a certain city or territory; which anniversaries fall into a certain year and which anniversaries to celebrate; it is possible to further refine the search in order to know which writers, poets, storytellers, philologists, filmmakers, artists, sportsmen, etc. are linked to a certain city or territory; it is also possible to carry out more targeted research by combining the various search criteria between them. It allows for answers to extremely targeted questions, such as:

- who are the illustrious people linked to the city of Palmanova and whose celebration is expected in 2018?
- what happened on the 3rd April?
- who are the illustrious women who made the Latisana city famous in the world?
- who are the Friulian athletes of the 20th century?
- who are the illustrious female Friulian poets?
- who are the illustrious people who have simultaneously performed the activities of writer, poet, painter and director?

There is no limit regarding the research which can be carried out. All this is possible thanks to the preliminary phase of finding the correct metadata to associate with each dictionary entry; this aspect has been the subject of a in-depth study to achieve the greatest flexibility in subsequent research phases (Fig. 2).

³ The printed version of the “Biographical Dictionary of Friulians. New Liruti” consists of three parts: “1. The Middle Ages”, published in 2006; “2. The Venetian Age”, published in 2009; “3. The Contemporary Age”, published in 2011, for a total of 7285 pages.

Fig. 2. The search form

3.6 Ability to Reach Users All over the World

Who will use the Biographical Dictionary of Friulians online? Certainly the Friulians, but let us not forget that the illustrious people in the dictionary are not only known in the region but throughout Italy and abroad as well. Let us not forget the thousands of Friulians in the world who will certainly appreciate such a tool; we think of scholars in various disciplines; in general let us think about anyone interested in a certain illustrious person and want deepen his biography online. In order to reach the widest visibility, special attention has been paid to the predisposition of all SEO (Search Engine Optimization) techniques aimed at getting the site to be the first result in searches made on various search engines. The aim is to let a user type for example “Caterina Percoto” and obtain as the first result the corresponding biographical entry on the dictionary. Particular importance will be given to verifying the number of accesses and pages viewed by users, using web analytics tools.

4 Technological Solutions Adopted

From a technical point of view, the key choices have been made on the basis of four guiding principles:

- adoption of open source technologies;
- using the most advanced and modern standards;
- making the resources available free of charge;
- independence from devices.

In particular:

- adoption of open technologies: the dictionary was developed using a nearly universal Content Management System (Wordpress). Although it is necessary constantly update the program, updates are freely available and programmers can access the source, therefore the widest guarantees against obsolescence is provided [12]. In other words, the work done will not become obsolete within a few years or even a few months (as is often the case when proprietary technologies are used). In fact, this it allows not to be bound by the company that has made the site since being an open technology, anyone in the future will be able to make changes to the dictionary or to further develop the work.
- use of the most advanced and modern standards: the infrastructure is based on today’s universally accepted standard technologies such as semantic web, RDF, facial navigation, etc. In addition, an endpoint was implemented through the sparQL query language. This means that anyone will have the ability to interface with the dictionary site to retrieve information of their own interest and export this information to their site (maximum openness). As far as we know this is the first case, if not all over the world at least in Italy, to apply these technologies to an online biography dictionary.
- resources available free of charge: access to biographical entries is available for everyone freely and free of charge; anyone can enjoy the work done and enrich their knowledge or even simply satisfy their curiosity by reading biographical entries of the dictionary available without geographical and temporal boundaries.
- independence from the device: the online edition makes content available in a richer and more accessible form by using a responsive technology that allows you to enjoy content using any type of device available today - not just computers but also tablets, phablets, smartphones and any other device, so that the dictionary can reach a much larger audience and regardless of the technology platform used.

5 Significant Technological Aspects

In the context of the world wide web, a semantic annotation provides information about the meaning of a resource and is intended to formally express its content, enabling it to be processable by machines [13]. Automatic resource annotation is an unresolved problem and usually involves human beings with the support of computer tools [16]. In this project, we have decided to make annotations to all biographical entries in the form of RDF triples, starting from information entered by a humanist team within the text (intext) and in external labels (meta tag). RDF triples, stored in a triplestore accessible from server resources exposed by the site, can also be queried with SPARQL, having made available a SPARQL endpoint [14]. In addition to representing one of the conditions required to make biographical entries and their content available as a linked data, the structure allows to carry out very complex queries (for example: what are the musicians who worked in the period 1820–1840 in the city of Aquileia? or: who are the

illustrious people that the city of San Daniele del Friuli must celebrate in 2018?). For example the query

```
PREFIX p:
SELECT DISTINCT ?postId ?sesso0 ?luogoNascita0
WHERE
{
?s p:postId ?postId .
?s p:sesso ?sesso0 .
?s p:luogoNascita ?luogoNascita0 .
FILTER ( ?sesso0 = "F" && ?luogoNascita0 = "Udine" )
}
```

retrieves the illustrious women born in Udine.

One of the objectives of the project was to identify a methodology that would allow the adding of semantic annotations to biographical entries by a team of humanists without geographical or temporal limitations and without the need to be a computer specialist. To achieve this goal, dedicated tools have been built on an easy-to-use and extremely popular Wordpress platform [17]. As a first step, the 2700 biographical entries, available in the PDF format used for the printed edition, were migrated in a hypertext format with automatic recognition of image positions and bibliographic references within the structure. The result was achieved by using an open source tool (pdf2html) that generated XML files with the appropriate formatting instructions; by elaborating such XML files, it was possible to highlight common patterns that allows us to associate them with the semantic aspect and thus rebuild the biographical entries in their original structure (title, subtitle, body, bibliography); it was also possible to extract the first external metadata (for example, the author of the biographical entry). The files thus produced have been read and imported into the database of the Wordpress platform, properly configured and customized. Subsequently, the 2620 biographical entries have been reviewed and annotated semantically using a specially developed tool. The tool that was used to facilitate the work of the review and annotation workgroup is an application that allows you to select and enter RDF element values with your mouse, limiting to the maximum the input of terms from the keyboard. A Wordpress plugin was then created to allow you to select the subject and indicate its property by choosing it from a drop down menu which lists the properties available in that context. The use was very easy and the simple and friendly graphic interface of Wordpress did the rest, allowing the team to get almost complete annotation in less time than originally expected, with great satisfaction from the workgroup. The metadata needed for semantic notation has been inserted in an extremely intuitive manner; in fact, it was enough to select the object element and assign it the appropriate tag (which represents the property) by choosing it from a context-sensitive drop-down list. A section has also been prepared for inserting the semantic annotation outside of the text, leaving editors the ability to enter metadata and postpone a post-text-revision phase to the creation of a controlled vocabulary for predicate

objects. This solution has been chosen to avoid the lengthy amount of time needed to create an internal ontology and to minimize the learning time of an external one. The bibliography has also been annotated and used to create RDF structures which describe external resources. The processed text is then dynamically read by a parser that creates RDF elements, which can take into account both text and external annotations. The terms to be inserted are based on a controlled lexicon and as far as this is currently local for the application, there is a configuration section (now accessible by the code) that will allow you to choose the ontology to use to represent the data outside and create correspondences between vocabulary and the internal one. The availability of a triplestore that can be queried by SPARQL has allowed to offer advanced search and navigation tools. It was possible to create pre-set searches for the user and a search form that proposes the properties and objects for the query, associating the requests with a SPARQL query and returning the responses to the user.

6 Future Developments

The online edition of the dictionary has also been an opportunity both for updating the biographical entries in the light of recent research and studies and to complete the work with the creation of a Supplement that is already under development: so new biographical entries currently missing (for various reasons, including the fact that in the paper edition there are no biographical entries of the characters who died after the printing date) will be added. The “engine” on which the site is based is extremely flexible and powerful and will allow further development in the future. For example, a feature that will be implemented will be geo-referencing the biographical entries, this will allow the user to “point and click” on a Friuli Venezia Giulia city and view all the illustrious people that are related to that city or to map geographically (on maps) the results of any search. Additional functionality can be implemented based on the feedback that will be received from users.

7 Conclusion

In conclusion, the Biographical Dictionary of Friulians (“New Liruti online”) has the ambitious aim of being not only the “digital version” of the printed version of the “New Liruti” but one of the richest and most structured deposits of cultural and historical information on the Italian web, based on the most innovative technologies, with the ability to reach a wider and potentially unlimited audience than the paper edition (consisting not only of scholars and researchers but also of students and ordinary people) so to become one of the most important cultural initiatives within the broadest project on “Cultural Identity of Friuli”.

References

1. Treccani - La cultura Italiana - Biografie. <http://www.treccani.it/biografie/>
2. Donne e uomini della Resistenza. <http://www.anpi.it/donne-e-uomini/>
3. Dizionario biografico dei protestanti in Italia. <http://www.studivaldesi.org/dizionario>
4. American National Biography Online. <http://www.anb.org>
5. ArchiDocWeb-RAH. <http://www.rah.es:8888>
6. Umetnosti, S.: Slovenska biografija. <http://www.slovenska-biografija.si>
7. Scalon, C., Griggio, C., Rozzo, U., Bergamini, G.: Nuovo Liruti. Editrice Forum, Udine (2011)
8. Greenstein, S., Zhu, F.: Do Experts or Collective Intelligence Write with More Bias? Evidence from Encyclopædia Britannica and Wikipedia. Harvard Business School (2014)
9. Documentation for ISO 639 identifier. <http://www-01.sil.org/iso639-3/documentation.asp?id=fur>
10. Google Books Ngram Viewer. https://books.google.com/ngrams/graph?content=Friulan%2C+Friulian&year_start=1800&year_end=2000&corpus=15&smoothing=3&direct_url=t1%3B%2CFriulan%3B%2Cc0%3B.t1%3B%2CFriulan%3B%2Cc0
11. Dizionario Biografico dei Friulani. <http://www.dizionariobiograficodeifriulani.it/?s>
12. Corrado, E.M.: The Importance of Open Access, Open Source, and Open Standards for Libraries. *Issues in Science and Technology Librarianship* (42) (2005). ISSN 1092–1206
13. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. *Sci. Am.* **284**(5), 28–37 (2001)
14. dizionario Biografico dei Friulani - SPARQL End Point. <http://www.dizionariobiograficodeifriulani.it/sparql/>
15. Reinert, M., Schrott, M., Ebneith, B.: Team deutsche-biographie.de: from biographies to data curation – the making of www.deutsche-biographie.de (2005). <http://ceur-ws.org/Vol-1399/paper3.pdf>
16. Tosi, D., Morasca, S.: Supporting the semi-automatic semantic annotation of web services: a systematic literature review. *Inf. Softw. Technol.* **61**, 16–32 (2015)
17. Munford, M.: How wordpress ate the internet in 2016... and the world in 2017, *Forbes* (2016). <https://www.forbes.com/sites/montymunford/2016/12/22/how-wordpress-ate-the-internet-in-2016-and-the-world-in-2017/#6926823d199d>

ISS Project: The Integrated Search System in the National Bibliographic Services

Luigi Cerullo^(✉)

Istituto Centrale per il Catalogo Unico delle biblioteche italiane e per le informazioni bibliografiche (ICCU), 00185 Rome, RM, Italy
luigi.cerullo@beniculturali.it

Abstract. The project originates from the need to overcome the criticalities of separate search and retrieval platforms for each of the national systems provided by the ICCU and at the same time to “heal the rift” between the National Union Catalogue (SBN) and the Digital Library (Internet Culturale). The actions aimed at rationalizing the retrieval model of ICCU systems, ensuring a consistent recall of information objects through its access interfaces, go towards two development lines: (1) use of one software platform that is the base for the application of the new Integrated Search System (ISS); this solution allows the creation of a single access point configured as a General Catalogue and, at the same time, the decommissioning of different OPACs whose peculiarities will be implemented in the single platform through dedicated search indexes; (2) native integration between bibliographic records and digital copies of publications through the integration of digital library system’s services and the SBN library management software.

Keywords: Collective catalogue · Digital library · Access interface
Management metadata · REST services · Searching and retrieval methods

1 Single Access and Retrieval Platform

1.1 General Goals of the Project

The main goal of the SRI project is to build a distributed information architecture that allows the creation of a single access point to the resources described in the main “management” databases, whose care depends directly and indirectly on ICCU (Edit16 [1], Manus On Line [2], The digital resource aggregator-index [3] and SBN [4]). This access interface or access point must be based on a retrieval model that takes into account the ontological level of the bibliographic universe (resources and related entities that are the subject of different descriptions or information items stored in the various management data-bases). The ontological coherence of the representation can be ensured only by acting directly in the bibliographic information creation phase (in the record making phase), ensuring the presence of identifiers and linking-keys between the different information objects - objects that are different for both data quality and data model. In fact, these objects are created in the specialist databases (Edit16 and Manus), are stored in the digital resource index (Internet Culturale) and are shared in the

Collective Catalogue (SBN), but often refer to the same resources and entities. The actions aimed at rationalizing the retrieval model of ICCU systems are listed below:

- (a) The use of a single software platform that is the basis of the new Integrated Search System (ISS) that allows the creation of a single access point configured as a general catalogue, and the decommissioning of the current infrastructure of separate OPACs, whose specific functionalities will be replicated in the single platform through dedicated search indexes.
- (b) Native integration between bibliographic records and digital copies of publications through the integration of digital library system's services with the most used SBN library management software that enables the SBN local nodes to supply the Index of digital resources (Internet Culturale) through automated generation of management metadata.

1.2 Integrated Search System and General Catalogue

The ISS is, at the application level, a single platform that includes a unified-index, meant as unified search profile. This solution allows a configuration of (a) search and retrieval interface (search interface, results lists and contextual search filters) of a single access point, presented as a general catalogue, and (b) individual search databases represented by the dedicated indexes Edit16, Manus and Internet Culturale, all these systems accessible through dedicated retrieval interfaces. The index of the General Catalogue is represented by the convergence of the SBN Unimarc profile and the Manus TEI-MS [5] profile, enriched by information of the digital copies described by the MAG [6] metadata referring to the analogic resource that has been described in SBN, in Manus or in Edit16 (1.3 sub) (Fig. 1).

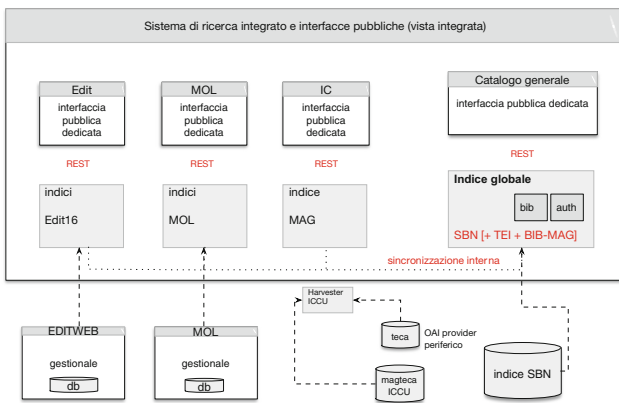


Fig. 1. In the scenery described, public retrieval interfaces represent the specialized sections (dedicated page-types) of a single web platform of fruition (Frontend Web-Applications). Specific search paths - in support of each specialized interface - will be included in the same database search server. The described layout provides that SBN's database functions as a basis for the unified profile.

Diagram of logical relationships between search database records. The diagram in Fig. 2 illustrates the relationships guaranteed in the indexing process of the unified-index.

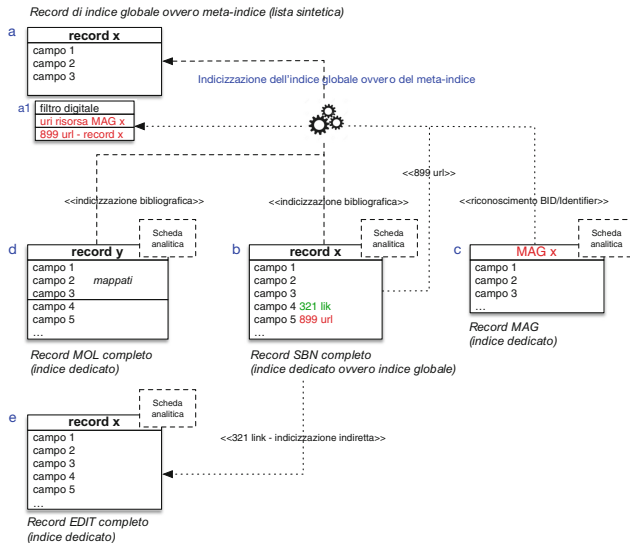


Fig. 2. (a+a1) lists the set of fields of the indices that configure the search masks and the unified-index results lists. For a given record x (e.g. a SBN resource) represented in the unified-index result set, the recognition procedure of a BID that is described into *dc:identifier* tags of indexed MAG (c) will populate index values that serve as filter (a1). This index contextually detect the presence of tags 899\$. The index profile (d) is able to represent TEI-MS information as a whole. Only a subset of these indices will be mapped to the unified-index profile. The index profile (e) represents the EDIT16 specialist search base that is not involved in the unified-index indexing phase. Each record of EDIT16 (here the bibliographic set is exemplified but the logic model is also valid for other entities related to the bibliographic record) is systematically invoked by links in the SBN index records (Unimarc tag 321 as a link to specialized repertories) created in the record-making phase. The index profile (b) represents the set of fields able to represent all SBN's Unimarc information. This information is the core of the integrated representation.

Re-engineering of specialized systems Edit16 and Manus as Client SBN-Marc. The cluster of records (Fig. 2) is made possible by two lines of development: (1) the re-engineering of the management data bases Edit16 and Manus-on-line configured as SBN clients (through the services of its application protocol SBNMarc [7]) and (2) the extension of functional architectures - intended to SBN Nodes - through which the integrated representation of digital resources (represented as bibliographic record attachments) is completed (1.3 sub). These environments will retain their specialist features and will be configured at the same time as Client SBNMarc (also thanks to the reuse of already available technical platforms) that can share all or part of the bibliographic information with the reference system (SBN), forming since the record-making phase the cluster of

distributed records, then coherently represented by the SRI retrieval machine (through its single access point).

Digital as a filter. In the architecture of the reengineered information system, based on the ISS, the digital index of Internet Culturale is not considered as a specialized data base. MAG records are largely de-structured representations of the same resources described in the SBN Catalogue or in Manus and Edit16. The configuration of the general catalogue indices would allow their registration as pure association with the records of the General Catalogue through the development of a unique ID recognition procedure that would come into play during the unified-index indexing phase. This architecture aim to populate specific indexes in order to preserve information about digital content, determined both by the presence of URIs associated in the field of SBN localizations and by the presence of the MAG record referring the same resource (compare diagram of logical relationships in Fig. 2).

1.3 Completion of Integrated Representation: Application Architecture for Peripheral Systems

The application architecture, planned for SBN local nodes, provides full integration of a digital library system's services with most used SBN library management software, through the development of batch generation services of management metadata, based on validated ICCU's mappings and wired in proposed systems. The designed architecture would allow to keep the records in the bibliographic catalogue (in shared environment) aligned with the management metadata (exhibited by the OI-Provider, component integrated in the system) and allow to describe the subset of records with digital attachments resident in the integrated digital library system. The MAG data-set or METS [8] data-set generated on the basis of batch procedures would be overwritten periodically.

Alignment over time. The functional diagram shown in Fig. 3, explains the solution destined to peripheral systems, together with the perspective of the development of an interface of access and retrieve with integrated unified-index: it will be developed in a long term plan in order to implement the integration-as-alignment between digital objects (represented by MAG or METS data-sets exhibited through OAI-PMH [9] Provider Systems) and bibliographic records of the collective catalogue (SBN), referring to the same resources. This alignment, guaranteed through overwrite cycles and exposure of management metadata, guaranteed on the basis of the implemented logic, would be represented in the central system (i.e. in the ISS, at the unified-index level) by the systematic retrieval of the reference to the digital object present in the single Index of digital resources.

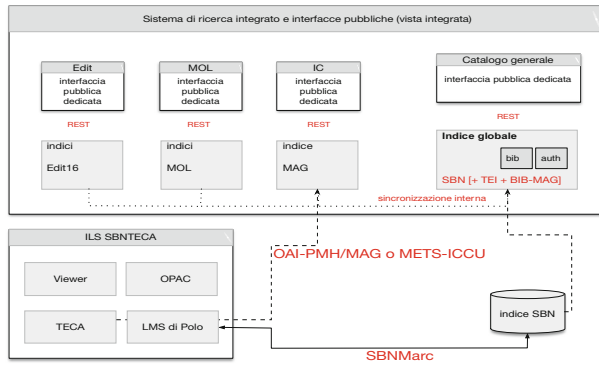


Fig. 3. Architecture of SBN nodes and integration with ISS. In this scenario, through an integrated management environment, the SBN nodes would also become content providers of Internet Cultural, represented - in the Integrated Search System - by one of its indexes: what we define in this work is a single index of digital resources.

The ISS unified-index is represented by the integration of the Unimarc profile of the SBN Index and the TEI-MS profile of the Manus system, enriched by information on the presence of digital copies described by MAG or METS records referring to the same resources described in SBN Index or in Manus and Edit16 data-bases. At first - based on the proposed architecture - only a subset of records that are present in the Single Digital Index (formerly the Internet Culture) would be retrieved, because they relate to digitized resources, mapped - through identifiers - with documents described in the Collective Catalogue. Over time - thanks to the diffusion of the functional architectures proposed in this chapter - the gap between the resources described in SBN catalogue (or within specialized systems) with digital attachments and that can be mapped with digital resources simultaneously described within the Single Digital Index, would gradually decrease.

1.4 Frontend Web Applications (Portals and Search Interfaces)

The Integrated Search System is represented, at the application level, by a single platform that communicates through REST services with Frontend web-applications that represent concretely the public access and return interfaces that the user can use (Fig. 4). These interfaces can be installed either in a single CMS, through which it will be possible to create the Portal of the General Catalogue, both in multiple instances of the same CMS. Each of them represent the service site of each specialized project (Edit16, Manus On Line, SBN, Internet Culturale). The main portal and portals dedicated to specialized databases would share the same template in order to maintain a “family area” according to the models already adopted in the most European projects.

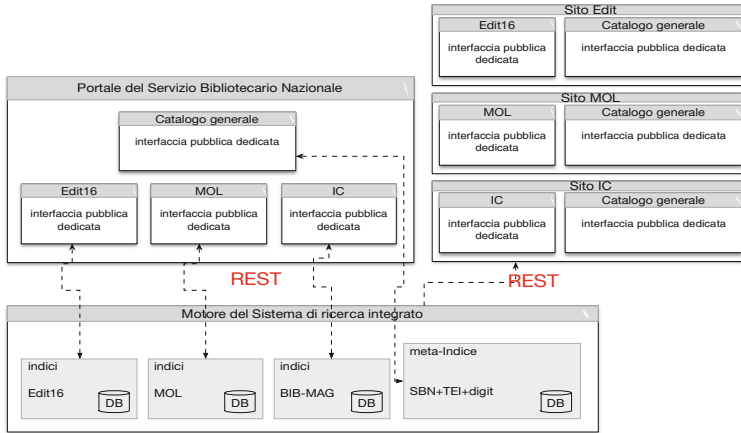


Fig. 4. Dynamic channels, represented by search maps, result-sets equipped with filtering and sorting tools, etc., are considered as “completed” functional components that work as plug-ins installed in CMS, on which the communication portals (or accompanying sites) are based. Avoid using the word “OPAC” to avoid confusion with self-consistent software, at the application level (with its own DB and administration area). The conceptual design of the application platform, based on the separation between Backend and Frontend model - communicating through REST services -, “reduces” the Frontend of the Integrated Search System to a set of components (4 components: Search Interface of General catalogue as a unified-index based on the SBN bibliographic profile, Edit16 Search Interface, Manus Search Interface, and Search Interface of MAG Index), which are deployed on multiple platforms, in different combinations depending on the use case.

References

1. Edit16, catalogue of italian editions of the 16th century. http://edit16.iccu.sbn.it/web_iccu/ihome.htm
2. Manus On Line (MOL), census of manuscripts held by Italian libraries. <https://manus.iccu.sbn.it/>
3. Internet Culturale, national digital resource aggregator. <http://www.internetculturale.it/opencms/opencms/it/>
4. Servizio Bibliotecario Nazionale, collective catalogue powered by peripheral catalogs through the services of the SBNMarc application protocol. <http://www.iccu.sbn.it/opencms/opencms/it/main/sbn/index.html?sessionId=E8E3A486F0E3C1F1A7F950414700E177>
5. TEI-MS Profile of the Manus On Line System. <https://jtei.revues.org/1054>
6. MAG Schema. <http://www.iccu.sbn.it/opencms/opencms/documenti/mag2-2006.html>
7. Technical Documentation and Schema of the SBNMarc Application Protocol. http://www.iccu.sbn.it/opencms/opencms/it/main/sbn/evoluz_indice_sbn/pagina_143.html
8. Documentation on METS metadata framework. <http://www.loc.gov/standards/mets/>
9. OAI-PMH schema and protocol. <https://www.openarchives.org/pmh/>

User Requirements and Relational Modelling for a Non-theatrical Cinema and Video-Art Cataloguing System

Petra Marlazzi¹, Lisa Parolo¹, Cosetta Saba¹, and Nicola Vitacolonna²(✉)

¹ La Camera Ottica Laboratory, Department of Humanistic Studies
and Cultural Heritage, University of Udine,
piazza Vittoria 41, 35170 Gorizia, GO, Italy

{[petra.marlazzi](mailto:petra.marlazzi@uniud.it),[lisa.parolo](mailto:lisa.parolo@uniud.it),[cosetta.saba](mailto:cosetta.saba@uniud.it)}@uniud.it

² Department of Mathematics, Computer Science and Physics,
University of Udine, via delle Scienze 206, 33100 Udine, UD, Italy
nicola.vitacolonna@uniud.it

Abstract. We describe an ongoing effort to design and implement a computerized cataloguing system for a laboratory dedicated to the restoration and archiving of non-theatrical cinema and video art. The goal is to evolve the current information system taking into account three different aspects: (i) national and international standards and workflows concerning preservation, cataloguing and archiving of film and contemporary art; (ii) specific needs emerging by daily experimentation in film and video restoration practice; (iii) the interoperability with film archives and contemporary art museums. A flexible conceptual Relational model based on Codd's RM/T is proposed as a first step towards the development of a system meeting the unique requirements of non-feature films and videos.

1 Introduction

The preservation, digitization and restoration of non-theatrical cinema and video art means, from a technical perspective, having to do with two moving-image carriers which differ, primarily, for the material they are made of, i.e., photo-chemical vs. electromagnetic. Both carriers are time-based, but while the former is 'isomorphic', human-readable and can be analysed without the help of a machine, the latter is 'polymorphic', machine-readable, and its content can only be checked using a compatible video-player. Moreover, even the same kind of carrier presents different physical formats (film: 8 mm, super8; video: 1/2" open reel, U-Matic, etc. . .) and varies in ontological¹ essence (e.g., theatrical or non-theatrical cinema, artwork or documentation)—which must be taken into

The paper has been discussed, planned and organized on the whole by the four authors. Petra Marlazzi has written Sect. 2 and Lisa Parolo has written Sect. 3.

¹ The term is used here with its philosophical meaning.

account when experimenting and practising preservation and digitization workflows [2, 3, 8–11, 13, 17, 19, 21, 22].

Once digitized, the specificity of the carrier (film or video) is partially lost. Digital and long-term preservation workflows are similar in a digital environment, and so is the cataloguing description. The ontological essence, though, must be preserved with minimal information loss, which is typically much more difficult for A/D conversion than for ‘digital natives’, hence requiring different protocols. The methodologies applied to obtain a digital version from an analogue source, too, must be carefully documented.

While digitization protocols have been widely discussed in the literature [4, 8–11, 13, 14, 17, 19, 21, 22], the cataloguing of non-theatrical cinema and video-art has so far received less attention. At the end of its activities our laboratory *La Camera Ottica* should be able to provide the institutions an interoperable database, keeping track of the salient phases that led to the creation of digital files. It is, then, always more necessary to build a database able to collect and organize all the preservative, administrative and historical/para-textual documentation [4, 6, 14–16].

There is currently a large number of standardized structures that can be used to start cataloguing cinema and contemporary art. Nevertheless, none of them is focused on the complexity and unique features of ‘objects’ such as non-theatrical cinema or video/time-based artworks, which can be considered at the boundary between cinema and art, thus requiring a multidisciplinary approach.

Most of the public Italian archives and museum institutions follow the ministerial OAC² (*Opere d’Arte Contemporanea*) regulation for cataloguing contemporary (and complex) artworks, which is a strictly hierarchical and very detailed collection of fields and subfields. Beside the lack of a high-level conceptual view of the data, the OAC specification presents many inadequacies in relation to the specific needs of ‘moving images’. For what concerns theatrical and non-theatrical cinema, a national regulation is missing and Italian film archives usually refer to the FIAF Cataloguing Manual³, which shares the same conceptual starting model as the FRBR⁴ (*Functional Requirements for Bibliographic Records*) report and adopts many of the definitions in the EN 15744 and EN 15907 European Standards for cinematographic works.⁵ As such, the FIAF reference does not contain recommendations focused on analogue videoart or, more generally, on time-based art. For what concerns art, most public European museums and archive institutions currently follow the guidelines of the DCA (*Digital Contemporary Art*) European project,⁶ which embraces the same conceptual model as FIAF and FRBR, but also includes suggestions on digitization workflows.

Most current proposals recognise, more or less explicitly, the need to consider the cataloguing process at different levels of detail. Some of them (FRBR,

² <http://www.iccd.beniculturali.it>.

³ <https://www.fiafnet.org/>.

⁴ <https://www.ifla.org/>.

⁵ <https://www.en-standard.eu>.

⁶ <http://www.dca-project.eu>. See, in particular, DCA dossiers D3.1, D4.2, and D6.1.

EN 15907, FIAF) distinguish, in slightly different ways, among four main ‘entities’ (see Table 1), whose meaning can be briefly summarised as follows:

Work: An entity comprising the intellectual or artistic content and the process of its realisation.

Expression/Variant: An entity that may be used to indicate any change to content-related characteristics that do not significantly change the overall content of a Work as a whole.

Manifestation: A physical embodiment of a moving image Work/Variant. Manifestations include all analogue, digital and online media.

Item: A single physical copy of a Manifestation of a Work or Variant.

A restoration laboratory deals with such entities in a bottom-up fashion, starting from physical artifacts typically taking the form of collections of items. In addition to the above entities, for a video and film preservation laboratory, the need to identify and catalogue the *collection in process* must be recognised, which warrants the creation of a further entity/level (Table 1).

Table 1. Comparison of conceptual entities for modelling moving image works.

FRBR	DCA	FIAF	OAC	La Camera Ottica
-	-	-	-	Collection
Work	Work	Work	Main record	Work
Expression	[Expression] ^a	[Variant] ^b	Level 1 record	Expression ^c
Manifestation	Manifestation	Manifestation	Level 2 record	Manifestation
Item	Item	Item	Level 3 record	Item

^a Mentioned but not used.

^b Optional.

^c An associative entity (see Sect. 4).

2 Specific Non-theatrical Cinema Requirements

What distinguishes a non-theatrical work from a feature movie is the purpose for which it is produced. The term “theatrical” is connected with the place where the official cinematographic production is shown and connotes a particular productive process—generally speaking, a kind of gauge (from 35 mm upward), specific production machines with professional figures, and complex copyright laws. Therefore, the term “non-theatrical” means every audiovisual work where screenings often take place within a private institutional context (e.g., film club screenings, educational screenings). Non-theatrical cinema peaked until the arrival of electronic and, later, digital technology which changed the carrier, though not the purpose, as is revealed by the categories of such films: industrial, training, scientific, amateur, ethno-anthropological, advertising and experimental.

The transition to digital, for some scholars, meant the “death of cinematography”, but, paradoxically, in the field of restoration and preservation it gave the possibility to show and watch vintage films again, especially those belonging to the often difficult to access non-theatrical heritage. This has brought to the attention of cultural institutions the necessity to preserve this inherently marginal cinematographic production. Moreover, the need for establishing specific guidelines to do that becomes even more compelling.

While the bibliographic models are suited for the needs of library catalogues, they are inadequate⁷ for moving image (non-feature film) archives, as the latter typically have in their collections rare or unique copies of films, or intermediate, and thus incomplete, productions and unreleased material. Even if the cataloguing standards provided by the FIAF take into consideration the “technological advances revolutionising cataloguing, preservation and access practices”,⁸ they do so from a perspective that values mainstream productions and thus only released materials. Those that are not are largely lacking of guidelines—an exception being the broadcast industry (for which broadcast-specific metadata schemas exist like EBUCore and PBCore). There is a need, then, for cataloguing strategies that do not rely on existing publishing requisites.

That is because an ‘archetype’ is a vacuous or difficult to define concept in the creative process of cinematographic production, whose existence (to paraphrase Walter Benjamin) is inseparable from its mechanical reproduction and its material properties, which are intrinsic to its essence [11]. Thus, in addition to a reconceptualization of the hierarchical model usually followed by textual criticism (see Table 1), the other particular characteristics that accompany the cataloguing of these materials are:

- the concurrence between *work* and *item*, since the filmic object that identifies the work is unique, in the sense that it does not have any copies to begin with;
- untitled work, resulting from the absence of a publishing intent or other artistic considerations;
- significance not as a work unto itself, but in relation to the collection of origin and to other *items* contained in it. To that end, there is a need for a cataloguing expansion supporting extended relationships and, in particular, para-textual documentation.

The publishing aspect being lacking, if not outright non-existing, it is no longer necessary to adopt a philological approach to recover the text and its classification. As a result, the concept of ‘expression/variant’ (Sect. 1) becomes less relevant for non-theatrical work, since the (compensative, substitutive, dismissal, alternative) variants of a work [21] appear only in a second edition—being variations of the original text—but non-theatrical work does not have an official publication and, accordingly, it does not have an ‘original’ text. Such characteristics point our attention to the ‘materiality’ of cinema rather than its

⁷ “While this shared bibliographic model works well for libraries, since many will have exact copies of the same publication, it does not provide all the functions that moving image archives need”, *The FIAF moving image cataloguing manual*, 2016, p. 2.

⁸ *The FIAF moving image cataloguing manual*, 2016, p. 2.

‘textuality’, highlighting the evolution of the techniques and functions of the cinematographic medium.

3 Specific Video Artworks Requirements

As mentioned in Sect. 1, video artworks are partially or completely recorded using magnetic tapes. Such support has widely varying characteristics, which depend, among the rest, on the storage format, which can be open-reel or cassette (U-Matic and VCR), on the variety of different brands (mostly Philips, Memorex and Sony) and standards used for recording (PAL, NTSC, SECAM), and on the age of the tape. The life of magnetic tapes is variable, but it can be estimated in about thirty years [3, 15, 16, 19], which is a relatively fast decay time.

Today’s best and most internationally agreed way for accessing and storing this particular type of artwork is to digitize the contents of the tapes. In general, though, digitization can only be performed after a technical restoration process of the analogue media and a historical contextualization of its contents through the analysis of para-textual documentation. Digitization then enables the production of copies using different coding formats, according to the foreseen purpose of each copy, and also the non-trivial task of long-term preservation storage (Hard Disk, LTO, Cloud) [20]. As it happens with film, this entails the current and future proliferation of analogue and digital copies, versions and variants, whose ‘authenticity’ and inter-relationships we should always be able to verify and re-establish. Each single aspect of the physical (diagnostic and historical) analysis within which to proceed to the first hypotheses regarding the state of preservation and contents of the tape must be identified and recorded in a catalogue system. While in the latest OAC regulation there is no indication concerning, in particular, the ‘material’ description of audiovisual tapes, some essential fields that describe time-based analog and digital carriers can be found in DCA dossiers [13, 16].

Moreover, both the DCA guidelines and the OAC regulation prescribe that technical interventions should be documented through both a final and more generic report and so-called *Preservation Metadata*. The OAC document, in particular, provides a paragraph (Conservation and Intervention) with data directed to diagnostics and restoration activities. However, although some parts of an intervention may be generalizable for all the elements of a single collection, other aspects, such as the empirical evaluation of each tape or the cleaning and baking processes, are defined *ad hoc*. A crucial feature of a cataloguing system, then, is to provide support for defining and recording *workflows*, which allows users to check the condition and validity of each phase of a restoration work, even in the long term [13, 16, 20].

Before moving from analog to digital signal, a fundamental aspect is the definition of the digitization quality parameters, which depend on the purpose for which each digital copy is created (conservation, access, preservation), on the physical characteristics of the analogue material, and on the available instrumentation. As a rule, the product of the migration is an *archival master file*; once the

content is verified through technical and historical analysis (which may bring to the need of digital restoration), the next intervention consists in the creation of a *production master file* and (possibly several) *derivative files* (also called *access copies*). Production and derivative files may be the result of post-production activities, such as editing, color correction, digital restoration, addition of titles, A/V encoding, and so on, each of which must be tracked in order to be verifiable and reversible [3, 16]. The compression parameters, in particular, must be carefully chosen according to the foreseen purpose of each copy, pondering the trade-off between quality and storage space (which ranges from a few to several hundreds GBs per copy).

The completion of the video preservation tasks is not the end of the cataloguing process of an artwork. As the DCA guidelines explain, the ultimate digital object to be catalogued will be composed of four fundamental parts: the work of art (abstract), its digital and analogue manifestations, documentation (layout plans, certificates and contracts) and contextual information (actors, locations, event, dates, etc. . .). Such parts can be linked to each other through administrative metadata; besides, each part requires its own specific descriptive metadata.

The structure proposed by OAC is deficient especially if the need is to describe genealogies between analogue and digital versions, variants and copies, the original and digital carrier and format in which audiovisual components are stored [12, 13, 15, 18]. OAC proposes a stratification distinguishing just between ‘main records’ (*scheda madre*) and ‘secondary records’ (*schede figlie*), that is, the components and under components of a complex artwork (see Table 1). Relationships between components are defined in a group of fields named ‘Relation’. The DCA research project, on the other hand, proposes three ‘closed’ hierarchical levels—*Work*, *Manifestation* and *Item*—sharing in this way the conceptual model of the FIAF cataloguing manuals (see Table 1). DCA guidelines, however, recognizing the uniqueness of each artwork, do not retain the *Expression/Variant* level; rather, they emphasise the description of the relations between two artworks that might share the same items and manifestations.

4 Summary of RM/T

We believe that building a comprehensive cataloguing system with future interoperability in mind requires first of all a rigorous conceptual and extendible view of the data. To that aim, we propose a formal specification and classification of the core entities described in the previous sections based on RM/T, which was developed as a more sophisticated version of the basic Relational Model for advanced users in need to model complex domains [5]. In what follows, we assume familiarity with the basic Relational Model (RM) [1].

RM/T, similarly to RM, includes not only the definition of the data structures, but also a rich variety of operators on those structures, and a number of integrity rules, without committing to a particular implementation. Besides, it supports incomplete descriptions (by having only a subset of properties defined

for an entity) and offers extended support for several semantic concepts. The explicit support at the conceptual level and a set of new powerful algebraic operators is what sets RM/T apart from RM and makes it suitable for modelling complex data management requirements.

The fundamental conceptual construct in RM/T is the *entity*. Informally speaking, an entity is any object, relationship or concept in the real world that has a relevant role for the information system to be built. Entities can be grouped into *entity types* (or simply *types*) via a form of abstraction from instances to classes (e.g., all persons may belong to a *Person* type).

Entities (and their types) can be partitioned, according to their role, into: *characteristic entities*, whose purpose is to describe multi-valued properties of entities of other types; *associative entities*, which denote relationships among other entities, and *kernel entities*, which are none of the above. Along an orthogonal dimension, entities (and their types) may be organized into taxonomies via set-inclusion based generalization (*subtyping*), e.g., every *Person is an Agent*. Types at the root of a taxonomy are called *inner types*. Finally, entities may be perceived at different levels of granularity: an entity may be viewed as a whole or as an aggregation of other (simple or aggregate) entities, e.g., a collection of movies. When an entity (type) represents an aggregation of other entities, it is called a *cover (type)*. Note that an entity in general may be a member of more than one cover (even of the same type).

Entities (of any type) are modelled using a countable domain \mathcal{ED} , whose elements are called *surrogates*, each one being a representative of one and only one distinct entity of the modelled reality. Two surrogates anywhere in the database are equal if and only if they denote the same real-world entity.⁹ For each type T (irrespective of its role), a unary relation on \mathcal{ED} , called an *E-relation*, is defined, which asserts the existence of an entity of type T . Besides, for each characteristic type C , an additional binary relation on $\mathcal{ED} \times \mathcal{ED}$ is defined, binding each entity of type C to the entity it describes; and for each associative type establishing a relationship among n other (not necessarily distinct) types, an $n + 1$ relation on $\mathcal{ED} \times \dots \times \mathcal{ED}$ is defined, in which one attribute identifies the associative entity and the remaining n attributes refer to the other interrelated entities.

Entities typically also have simple (immediate) properties that describe them; in particular, in most cases they (should) have simple properties *identifying* them, i.e., subsets of properties assuming a unique value for each distinct entity. Simple properties are modelled using *property relations (P-relations)*, i.e., n -ary relations $P(S : \mathcal{ED}, A_1 : D_1, \dots, A_n : D_n)$, where S identifies (to the system) the entity being described, and each attribute A_i is defined on a suitable domain D_i (which constrains the admissible values for property A_i). If $D_i = \mathcal{ED}$ for some i then P is called a *designative* relation. We adopt the view that property relations should be decomposed into *minimal meaningful units* [5]: normalization theory can be used to determine such groupings.

⁹ Of course, such bijection can be enforced by the system only as long as natural keys can be defined for the entities involved—RM/T does not require them, though.

The characteristic types providing a description of a given kernel or associative type T form a strict hierarchy (the *characteristic tree* of T). So, an RM/T model is essentially a collection of characteristic trees, whose nodes are further connected by many-to-many relationships (via associative relations) or one-to-many relationships (through designative relations). We will further assume that, within a single characteristic tree, all non-surrogate attributes and all surrogate attributes not referring to the same type have distinct names.

RM/T also maintains explicit meta-information about the database schema in a collection of *catalog relations* (see Fig. 3), which includes the following: relation **PG** ties each P -relation to its E -relation; relation **CG** relates each characteristic type to the type it describes; relation **AG** stores the fact that a type *sub* is part of the definition of an associative type *sup* via attribute *att*; relation **SG** describes the immediate subtypes *sub* of each generic type *sup*; and relation **KG** specifies which types *sub* may be members of cover types *sup*.

RM/T enforces a number of integrity constraints in addition to those that are part of RM. Some of them should be obvious from the foregoing description (e.g., a tuple t may appear in a P -relation only if the corresponding E -relation asserts the existence of the entity described by t ; referential integrity on designative attributes; and so forth). One constraint that will be useful to keep in mind is that *every occurrence of a surrogate anywhere in the database must appear in at least one E-relation*. Another important constraint is that each characteristic entity is existent-dependent on the entity it refers to (which is not required for one-to-many relationships in general). See [5] for the full list of RM/T constraints.

Finally, one of the more interesting aspects of RM/T is its extended Relational Algebra, which allow users to formulate queries that are somewhat independent of the schema of a database. We consider the following operators:

1. $\text{NOTE}(R)$ is the name of the relation R (i.e., a string). The inverse operator is $\text{DENOTE}()$ (i.e., $\text{DENOTE}(\text{NOTE}(R)) = R$).
2. $\text{TAG}(R) \doteq R \times \{\text{NOTE}(R)\}$.
3. $\text{COMPRESS}(\cdot, \mathcal{R})$ is the relation obtained by repeated pairwise application of associative and commutative operator \cdot to the relations in the set \mathcal{R} ;
4. $\text{APPLY}(f, \mathcal{R}) \doteq \{f(r) \mid r \in \mathcal{R}\}$, where f maps relations into relations;
5. $\text{CLOSE}(R)$ is the transitive closure of (binary) relation R .
6. $\text{PROPERTY}(R)$ groups into a single relation E -relation R and all its immediate properties.¹⁰

We find it convenient to define additional operators $\text{GRAPH}()$ and $\text{LGRAPH}()$. Given the name of any associative E -relation A , $\text{GRAPH}(A)$ is a binary relation on schema $(s: \mathcal{ED}, t: \mathcal{ED})$ representing the symmetric closure of the graph of the association denoted by A . Then, $\text{LGRAPH}(A) \doteq \text{TAG}(\text{GRAPH}(A))$. For instance, if a , b and c are related via a ternary association A then $\text{LGRAPH}(A)$ contains the tuples (a, b, A) , (b, a, A) , (a, c, A) , (c, a, A) , (b, c, A) , and (c, b, A) .

¹⁰ This is a derived operator. See [5] for a formal definition.

5 A Case Study: *Do You Remember This Movie?*

To give an example of the complexity we face when dealing with particular kind of non-theatrical cinema and time-based artworks, we will consider *Do You Remember this Movie?* by Luigi Viola as a case study. In 1979, the video was recorded on an analogue U-Matic cassette. It depicted the artist while watching a home movie he had made with his family a couple of years before (Fig. 2). The institution owning the video had many copies of the same artwork in several different (VHS and DVD) cartridges. Some time after the acquisition by our lab, we found out that, in 1982, the artist had produced a ‘remake’ of the video with the same content and title, but with different production technologies. Besides, further documental research revealed that the first version of *Do You Remember this Movie?* had been featured as part of a multi-media installation entitled *I looked for... (da Alice 1977)*, which was presented during a collective exhibition in 1980 [7] (see Fig. 1). Eventually, we also found the film of the home movie which had been projected in both versions of *Do You Remember this Movie?*

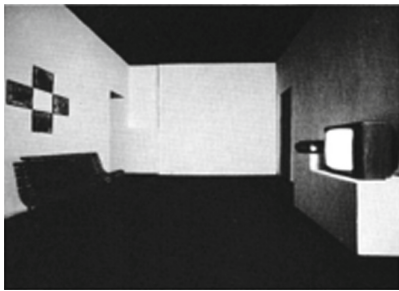


Fig. 1. The only known installation of *I Looked for... (da Alice 1977)*.



Fig. 2. A frame from *Do You Remember This Movie?* (1979).

We face, then, a set of complex relationships among: one ‘original’ artwork created in 1979; a different version with the same title dated 1982; a home-movie film (with no title) made between 1975 and 1976 by the same artist; and a multimedia installation with a temporal component constituted by four different videos, among which *Do You Remember this Movie?*, and a spatial component, including the room, a green bench and four photographs (see Fig. 1). Note also that the multimedia installation exists only through the para-textual documentation and the testimony of the artist.

“A good catalog is designed to demonstrate to catalog users a number of different kinds of relationships among its records.” [23]. Yet, the whole reconstruction of the preservation and exhibition history of the artwork is the most difficult part of cataloguing. Establishing complex relationships also makes retrieving data more difficult. We claim that RM/T can help users designing systems in which such problems are elegantly solved.

The core of our model is based on four inner kernel relations corresponding to the types **Collection**, **Item**, **Manifestation** and **Work**. The inextricable relationship between abstract works and their concrete manifestations is captured by treating **WORK** as a cover type having **MANIFESTATION** instances as members. According to such point of view, a work is essentially just a less granular view of a set of manifestations. Similarly, **MANIFESTATION** and **COLLECTION** are also cover (in fact, partition) types with **ITEM** instances as members.

The **Variant/Expression** type is modelled as an associative relation between works (possibly with additional constraints, such as the requirement that it is a strict partial order). Using an associative type circumvents the inherent semantic overlap between ‘expressions/variants’ and ‘works’, which would occur if **Variant** were treated as kernel. It also removes the transitive dependency (works have many variants or manifestations, and variants have many manifestations) implied by the **FIAP** proposal.

To account for mixed media entities such as *I Looked For...* and *Do You Remember This Movie?*, we must recognise that the former is *not* a manifestation of the latter, nor a variant; instead, the latter is an essential *constituent* element of the former. Mixed media items are *composed of* A/V items and extra-filmic material (e.g., photographs, objects, spaces). Making such composition relationships explicit is essential for cataloguing, because reuse is common in video art and non-theatrical cinema—in fact, we argue that **Composition** is at least as important as **Variant**. Compositions may be specified at any level of granularity (items, manifestations, works), hence they must be modelled using three (not independent) associative types: **ITEMCOMPOSE**, **MANIFCOMPOSE**, **WORKCOMPOSE**. For instance, *Do You Remember This Movie?* is a component of *I Looked For...* at the work level; the 1978’s version is a component of the (only) manifestation of *I Looked For...*; but, as it happens, we do not know any relationship at the item level.

Finally, for each defined entity type, an arbitrary number of properties and characteristics can be defined.¹¹ Many RM/T queries make no assumption on the number and structure of *P*-relations or on the size and depth of characteristic trees. We conclude this section by showing a few such queries.

1. Retrieve *all* the immediate properties of the original version of *Do You Remember This Movie?*:

$$\begin{aligned} W &\leftarrow \sigma_{\text{title}='Do\ You...'}(\text{PROPERTY}(\text{WORK}) \bowtie \text{PROPERTY}(\text{TITLE})) \\ V &\leftarrow \delta_{v\text{work}\# \rightarrow \text{work}\#}(\pi_{v\text{work}\#}(\text{VARIANTWORK})) \\ \text{Answer} &\leftarrow W \bowtie (\pi_{\text{work}\#}(\text{VARIANTWORK}) \setminus V) \end{aligned}$$

Here, δ is the *renaming* operator [1].

2. Retrieve the characteristic tree of 1982’s *Do You Remember This Movie?*:

$$\begin{aligned} R &\leftarrow \pi_{\text{sub}}(\sigma_{\text{sup}=\text{WORK}}(\text{CLOSE}(\text{CG}))) \cup \{\text{NOTE}(\text{WORK})\} \\ S &\leftarrow \text{COMPRESS}(\bowtie, \text{APPLY}(\text{PROPERTY}(), \text{APPLY}(\text{DENOTE}(), R))) \\ \text{Answer} &\leftarrow \sigma_{\text{title}='Do\ You\ Remember...'\wedge \text{year}=1982}(S) \end{aligned}$$

¹¹ For a (somewhat contrived) example, see Appendix A.

CATR			PG		CG		
<i>relname</i>	<i>reltype</i>		<i>sub</i>	<i>sup</i>	<i>sub</i>	<i>sup</i>	
COLLECTION	<i>E</i> -relation, inner kernel		WORKTYPE	WORK	WORKTITLE	WORK	
ITEM	<i>E</i> -relation, inner kernel		WORKYEAR	WORK	UNIT	ITEM	
ANALOGITEM	<i>E</i> -relation, kernel		VARIANTWORK	VARIANT	INTERVENTION	UNIT	
DIGITALITEM	<i>E</i> -relation, kernel		
MANIFESTATION	<i>E</i> -relation, inner kernel		KG		SG		
WORK	<i>E</i> -relation, inner kernel		<i>sub</i>	<i>sup</i>	<i>sub</i>	<i>sup</i>	
VARIANT	<i>E</i> -relation, associative		ITEM	COLLECTION	DIGITALITEM	ITEM	
WORKCOMPOSE	<i>E</i> -relation, associative		ITEM	MANIFESTATION	ANALOGITEM	ITEM	
MANIFCOMPOSE	<i>E</i> -relation, associative		MANIFESTATION	WORK	PERSON	AGENT	
ITEMCOMPOSE	<i>E</i> -relation, associative		ORGANISATION	AGENT	
WORKTYPE	<i>P</i> -relation		
WORKTITLE	characteristic relation		AG (cont.)				
AG			<i>sub</i>	<i>sup</i>	<i>att</i>		
...	...		SUBJECT	HASASUBJECT	<i>subject#</i>		
WORK	VARIANT	<i>work#</i>	WORK	HASASUBJECT	<i>work#</i>		
WORK	VARIANT	<i>vwork#</i>	WORK	CREDITS	<i>work#</i>		
WORK	WORKCOMPOSE	<i>work#</i>	AGENT	CREDITS	<i>agent#</i>		
WORK	WORKCOMPOSE	<i>complex_work#</i>		
...					

Fig. 3. (A small part of) the RM/T catalog.

The result is a flattened version of the characteristic tree of the given work.

3. Find the persons who are the subject of works they have authored:

$$C \rightarrow \sigma_{role='author'}(\text{PROPERTY}(\text{CREDITS}) \bowtie \text{PROPERTY}(\text{ROLE}))$$

$$\text{Answer} \rightarrow \delta_{subject\# \rightarrow agent\#}(\text{PROPERTY}(\text{HASASUBJECT})) \bowtie C$$

Since an entity may belong to several types, to assert that a person is the subject of a work it is sufficient to insert its surrogate into a kernel type SUBJECT. An associative type HASASUBJECT then may relate any subject with any work.

4. Find the works related *in any way* to 1982's *Do You Remember This Movie?*:

$$A \leftarrow \pi_{sup}(\sigma_{sub='WORK'}(\mathbf{AG} \bowtie_{sup=sup' \wedge sub=sub' \wedge att \neq att'} \delta_{* \rightarrow *}(\mathbf{AG})))$$

$$B \leftarrow \pi_{work\#}(\sigma_{title='Do You...'\wedge year=1982}(\text{WORKYEAR} \bowtie \text{PROPERTY}(\text{TITLE})))$$

$$\text{Answer} \leftarrow \delta_{work\# \rightarrow s}(B) \bowtie \text{COMPRESS}(\cup, \text{APPLY}(\text{LGRAPH}(), A))$$

This query returns tuples of the form (w_1, w_2, R) , where w_1 is the surrogate of the specified video, and w_2 is directly related to w_1 via relation R .

6 Concluding Remarks

None of the current national and international standards and regulations, while critical to ensure future interoperability with other databases and institutions, suit the specific needs for cataloguing complex 'objects' like non-theatrical cinema and video/time-based art. An accurate description of the complex relationships among the several entities involved in the restoration process and well-defined, system-supported, digitization and cataloguing workflows are two of the

key elements that we have identified as crucial for a useful cataloguing system ready for interoperability. We have proposed an extendible conceptual model as a foundation for such a system. Our model can be easily implemented in relational DBMSs (although support for RM/T is currently lacking) or mapped into metadata schemas.

A Example RM/T Instance for a Video Art Catalogue

Mandatory properties such as internal or standard identifiers (i.e., natural keys) are omitted for simplicity. Attributes ending with # are defined on \mathcal{ED} . Surrogate keys have been underlined only when there is more than one surrogate attribute.

A.1 Some Kernel Entities and their *P*-Relations

<u>COLLECTION</u>	<u>COLLECTIONNAME</u>		<u>ITEM</u>	<u>ITEMCOLLECTION</u> (cover member)		<u>ITEMMANIFESTATION</u> (cover member)	
<u>coll#</u>	<u>coll#</u>	<u>coll_name</u>	<u>item#</u>	<u>item#</u>	<u>coll#</u>	<u>item#</u>	<u>manif#</u>
<u>k₁</u>	k ₁	Fondo Cavallino	i ₁	i ₁	k ₁	i ₁	m ₁
			i ₂	i ₂	k ₁	i ₂	m ₂
			i ₃	i ₃	k ₁	i ₃	m ₃
			i ₄	i ₄	k ₁	i ₄	m ₄

<u>ANALOGITEM</u>	<u>ANALOGDATA</u>		<u>DIGITALITEM</u>	<u>DIGITALDATA</u>	
<u>analog#</u>	<u>analog#</u>	<u>base extent</u>	<u>digital#</u>	<u>digital#</u>	<u>container</u>
i ₁	i ₁	triacetate 6474 ft	i ₄	i ₄	MPEG
i ₂					
i ₃					

<u>WORK</u>	<u>WORKTYPE</u>		<u>WORKYEAR</u>	<u>FORMAT</u>	<u>FORMATINFO</u>		
<u>work#</u>	<u>work#</u>	<u>work_type</u>	<u>work#</u>	<u>format#</u>	<u>format#</u>	<u>carrier</u>	<u>format</u>
w ₁	w ₁	Home Movie	w ₁	f ₁	f ₁	video	U-Matic
w ₂	w ₂	Video Art	w ₂	f ₂	f ₂	film	16 mm
w ₃	w ₃	Mixed Media	w ₃	f ₃	f ₃	video	H.264
w ₄	w ₄	Video Art	w ₄				

<u>MANIFESTATION</u>	<u>MANIFWORK</u> (cover member)		<u>MANIFFORMAT</u> (designative)	
<u>manif#</u>	<u>manif#</u>	<u>work#</u>	<u>manif#</u>	<u>format#</u>
m ₁	m ₁	w ₁	m ₁	f ₂
m ₂	m ₂	w ₂	m ₂	f ₁
m ₃	m ₃	w ₃	m ₄	f ₁
m ₄	m ₄	w ₄	m ₅	f ₃
m ₅	m ₅	w ₂		

<u>SUBJECT</u>	<u>SUBJECTDESCRIPTION</u>		<u>AGENT</u>	<u>AGENTADDRESS</u>	
<u>subject#</u>	<u>subject#</u>	<u>description</u>	<u>agent#</u>	<u>agent#</u>	<u>address</u>
s ₁	s ₁	Family	a ₁	a ₃	via della video arte 78
s ₂	s ₂	Carnival	a ₂	a ₅	via delle Scienze 205
w ₁			a ₃		
a ₁			a ₄		
			a ₅		

PERSON	PERSONNAME	ORGANISATION	ORGANISATIONNAME
<u>person#</u>	<u>person# first last</u>	<u>org#</u>	<u>org# org_name</u>
a ₁	a ₁ Luigi Viola	a ₄	a ₅ Galleria del Cavallino
a ₂	a ₂ Paolo Cardazzo	a ₅	a ₆ La Camera Ottica
a ₃	a ₃ Lisa Parolo		

A.2 Some Characteristic Entities and their *P*-Relations

TITLE	TITLEWORK	TITLEDDETAILS
<u>title#</u>	<u>title# work#</u>	<u>title# title title_type</u>
t ₁	t ₁ w ₂	t ₁ Do You Remember This Movie? preferred
t ₂	t ₂ w ₂	t ₂ Do You Remember This Film? draft
t ₃	t ₃ w ₃	t ₃ I Looked for... (da Alice 1977) preferred
t ₄	t ₄ w ₄	t ₄ Do You Remember This Movie? preferred

UNIT	UNITITEM	UNITDETAILS
<u>unit#</u>	<u>unit# item#</u>	<u>unit# unit_details</u>
u ₁	u ₁ i ₁	u ₁ reel 1
u ₂	u ₂ i ₁	u ₂ reel 2

INTERVENTION	INTERVENTIONUNIT	INTERVENTIONDESCRIPTION
<u>int#</u>	<u>int# unit#</u>	<u>int# intervention_description</u>
j ₁	j ₁ u ₁	j ₁ manual cleaning
j ₂	j ₂ u ₁	j ₂ scanning

OWNERSHIP	OWNERSHIPCOLL	OWNERSHIPOWNER	OWNERSHIPACQUISITION
<u>own#</u>	<u>own# coll#</u>	<u>own# agent#</u>	<u>own# acq_date</u>
o ₁	o ₁ k ₁	o ₁ a ₅	o ₁ 1976/1/1
o ₂	o ₂ k ₁	o ₂ a ₆	o ₂ 2003/1/1

PARATEXT	PARATEXTMANIFESTATION	PARATEXTDESCRIPTION
<u>para#</u>	<u>para# manif#</u>	<u>para# paratext_description</u>
p ₁	p ₁ m ₁	p ₁ see Fig. 1
p ₂	p ₂ m ₁	p ₂ catalogue

A.3 Some Associative Entities, with Properties and Characteristics

VARIANT	VARIANTWORK	VARIANTNOTES
<u>variant#</u>	<u>variant# work# vwork#</u>	<u>variant# variant_notes</u>
v ₁	v ₁ w ₂ w ₄	v ₁ Some parts remade

MANIFCOMPOSE	MANIFCOMPOSEMANIF
<u>mc#</u>	<u>mc# manif# complex_manif#</u>
e ₁	e ₁ m ₁ m ₂

WORKCOMPOSE	WORKCOMPOSEWORK
<u>wc#</u>	<u>wc# work# complex_work#</u>
e ₁	e ₁ w ₁ w ₂

<u>CREDITS</u>	<u>CREDITSAGENTWORK</u>			<u>HASASSUBJECT</u>	<u>WORKHASASSUBJECT</u>		
<u>credits#</u>	<u>credits#</u>	<u>agent#</u>	<u>work#</u>	<u>has#</u>	<u>has#</u>	<u>subject#</u>	<u>work#</u>
c_1	c_1	a_1	w_2	h_1	h_1	w_1	w_2
c_2	c_2	a_2	w_2	h_2	h_2	w_1	w_4
				h_3	h_3	a_1	w_2

<u>ROLE</u>	<u>ROLECREDITS</u>		<u>ROLEDESCRIPTION</u>	
<u>role#</u>	<u>role#</u>	<u>credits#</u>	<u>credits#</u>	<u>role</u>
r_1	r_1	c_1	r_1	author
r_2	r_2	c_1	r_2	producer

References

1. Abiteboul, S., Hull, R., Vianu, V.: Foundations of Databases. Addison Wesley, Reading (1995)
2. Bolter, J.D., Grusin, R., Grusin, R.A.: Remediation: Understanding New Media. MIT Press, Cambridge (2000)
3. Bordina, A.: La conservazione dell'arte video: teorie, strategie e tecniche. Ph.D. thesis, University of Udine (2008)
4. Carlos, M.: A comparison of Scanning Technologies for Archival Motion Picture Film. Master's thesis, Staatliche Akademie der Bildenden Künste Stuttgart (2013)
5. Codd, E.F.: Extending the database relational model to capture more meaning. ACM Trans. Database Syst. **4**(4), 397–434 (1979)
6. Ernst, W.: Digital Memory and the Archive. University of Minnesota Press, Minneapolis (2013)
7. Fagone, V. (ed.) Camere incantate. Video, cinema, fotografia e arte negli anni '70. Milano, Palazzo Reale (1980)
8. Flueckiger, B.: Material properties of historical film in the digital age. NECSUS—Eur. J. Media Stud. **1**(2), 135–153 (2012)
9. Fossati, G.: From Grain to Pixel: The Archival Life of Film in Transition, 2nd edn. Amsterdam University Press, Amsterdam (2011)
10. Gracy, K.F.: Documenting the process of film preservation. Moving Image **3**(1), 1–41 (2003)
11. Herr, L.: The digital dilemma: Strategic issues in archiving and accessing digital motion picture materials. Academy of Motion Picture Arts and Sciences (2007)
12. Laurenson, P.: Authenticity, change and loss in the conservation of time-based media installations. Tate Papers Autumn 2006 (2006)
13. Martone, P. (ed.) Tra memoria e oblio. Percorsi nella conservazione dell'arte contemporanea, I timoni (2014)
14. Mazzanti, N.: Challenges of the digital era for film heritage institutions. Publications Office of the European Union (2012)
15. Noordegraaf, J., Saba, C., Le Maître, B., Hediger, V.: Preserving and Exhibiting Media Art: Challenges and Perspectives. Amsterdam University Press, Amsterdam (2013)
16. Parolo, L.: For a History of Italian Video-Art in the Seventies: the Cavallino Gallery Video Archive (1970–1984). Historical-Critical Examination of Sources and Research of New Digital Cataloguing and Archiving Methods. Ph.D. thesis, University of Udine and University Roma 3 (2017)
17. Read, P., Meyer, M.P.: Restoration of Motion Picture Film, 1st edn. Butterworth-Heinemann, Oxford (2000)

18. van Saaze, V.: *Installation Art and the Museum: Presentation and Conservation of Changing Artworks*. Amsterdam University Press, Amsterdam (2013)
19. Saba, C.G.: *Arte in Videotape: Art/tapes/22, collezione ASAC - La Biennale di Venezia*. Conservazione, restauro, valorizzazione. Silvana editoriale (2007)
20. Traczyk, T., Ogryczak, W., Pałka, P., Śliwiński, T.: *Digital Preservation: Putting It to Work*. Springer, Heidelberg (2017)
21. Venturini, S.: *Il Restauro Cinematografico: Principi, Teorie, Metodi*, Campanotto Editore (2006)
22. Wheeler, J.: *Videotape Preservation Handbook*. AMIA (2002)
23. Yee, M.M.: *Moving Image Cataloging: How to Create and How to Use a Moving Image Catalog*. Libraries Unlimited (2007)

The European Project OpenUP: OPENing UP New Methods, Indicators and Tools for Peer Review, Impact Measurement and Dissemination of Research Results

Alessia Bardi, Vittore Casarosa^(✉), and Paolo Manghi

ISTI-CNR, Pisa, Italy

{alessia.bardi,vittore.casarosa,paolo.manghi}@isti.cnr.it

Abstract. Open Access and Open Scholarship are substantially changing the way scholarly artefacts are evaluated, published and assessed, while the introduction of new technologies and media in scientific workflows has changed the “how and to whom” science is communicated, and how stakeholders interact with the scientific community. OpenUP addresses key aspects and challenges of the currently transforming science landscape. Its main objectives are to: (i) identify and determine new mechanisms, processes and tools for the peer-review of all types of research results (publications, data, software, processes, etc.); (ii) explore, identify and classify innovative dissemination mechanisms with an outreach aim towards businesses and industry, education, and society as a whole; (iii) analyse and identify a set of novel indicators that assess the impact of research results and correlate them to channels of dissemination.

OpenUP is engaged with research communities from life sciences, social sciences, energy, arts and humanities, implementing a series of hands-on pilots to assess and verify the proposed new mechanisms for the cycle review-disseminate-assess, to understand how these mechanisms correspond to the requirements and needs of the research communities. The final outcome of the project will be a set of concrete, practical, validated policy recommendations and guidelines for all stakeholders, namely academia, industry and government institutions.

Keywords: Open access · Open science · Open scholarship · Peer review
Impact assessment

1 Objectives

Open Access, Open Science, Open Scholarship accompanied by sharing enabling technologies, have revolutionized the way scholarly artefacts are evaluated, published and assessed. These developments have also changed the requirements and practices of the involved stakeholders, namely researchers, publishers, funders, institutions, industry and the public. The exponentially growing research output, the increasing demand for a more open, transparent and reproducible science, as well as apparent shortcomings in present quality assurance and evaluation methods require key stakeholders to re-think the very nature of how the quality of research artefacts is evaluated. In addition, novel and innovative ways of disseminating research outputs revolutionise the ways how and

to whom science is communicated, and how stakeholders interact with the scientific community.

Traditional ways of publication and evaluation do not satisfy the needs of this changing landscape and currently there are more open questions than answers. How can we determine and ensure the quality level of research artefacts, if the standard evaluation methods are no longer useful? Which metrics can be used to evaluate new forms of publishing (data, software), which go beyond the traditional bibliometric used for books and papers? How do technological advancements and the integration of Open Science workflows and behaviours affect the new landscape? How do different stakeholders measure the impact of science? How do we adapt the policy framework so that it becomes more open and gender sensitive? How can we measure the impact of research findings on society and businesses outside the traditional evaluation and publishing channels? What are the new business and pricing models that need to be put in place?

The review-disseminate-assess cycle is a multifaceted process involving different stakeholders:

- Publishers, who have yet to understand and adapt to new reviewing methods, and still measure their success through bibliometric;
- Researchers, especially the young ones, who instinctively find novel ways to disseminate their research but are lacking a way to measure their success;
- Policy makers (e.g. funders), who strive to make evidence based assessments but do not have the tools to move beyond the current status quo;
- Institutions, who need to integrate new indicators for researcher career advancement, adapt to emerging business models for journal subscriptions, expand their services for data management, or assess their research outcome;
- Citizens and industry who use science and implicitly increase the scientific impact.

There are already many initiatives and projects addressing an “open peer review process”, or addressing new and different impact indicators, or experimenting innovative dissemination methods (see the Reference section for a selected bibliography). OpenUP intends to push forward these fields by addressing the key aspects and challenges of the currently transforming science landscape in terms of quality assurance, communication of scientific outputs, and impact assessment with a focus on Open Science developments. The main objectives of the project can be summarized as follows.

- *Explore, analyse and promote open peer review mechanisms.* Identify and determine novel mechanisms, processes and tools for peer-review for all types of research outcomes. Investigate and understand how these are adapted and applied in an Open Science, e-Infrastructure enabled environment. One of the relevant emerging trends is the requirement to save and assess the “Research Flow”, i.e. the process by which research results are produced by applying a certain methodology to certain data. OpenUP is studying how peer-review practices and methods can be applied, adapted and extended beyond articles, books and monographs to include research data, research flow and software.
- *Explore and promote innovative methods of research dissemination and communication.* Explore, identify and classify innovative dissemination mechanisms and their effectiveness, suitability and impact. Study communication mechanisms that go

beyond the traditional scientific academic venues with an outreach aim towards businesses and industry, education, and society as a whole.

- *Define research metrics and indicators for different stakeholders.* Collect a set of indicators that assess the impact of various types of research results in an open, social network savvy environment, and put them into perspective in terms of channels of dissemination. Investigate the commonalities and differences on how these are perceived, adapted and used by the various research communities and involved stakeholders.
- *Validate the OpenUP framework with community driven pilots.* Engage with research communities from life sciences, social sciences, energy, arts and humanities, and implement a series of hands-on pilots to assess and verify the proposed new mechanisms for the cycle review-disseminate-assess, to understand how these mechanisms correspond to the requirements and needs of the research communities.

2 Overall Approach

OpenUP is following a phased approach over its three main pillars of Review-Disseminate-Assess. These phases, namely Landscaping – Initial analysis – Assessment and validation – Policy review – Synthesis (see Fig. 1), will feed to and run in parallel to an intensive awareness and dissemination activities. All results from one phase will be fed into the next phases, while they will also be made public for consultation through the OpenUP’s platform.

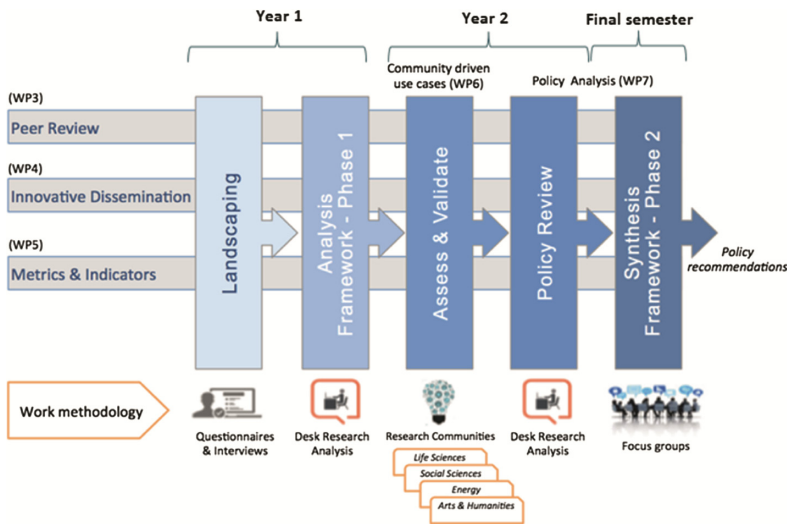


Fig. 1. OpenUP overall methodology

The project, started in 2016, has completed the phases for Year 1 (Landscaping and Analysis Framework – Phase 1) and just started the Assess & Validate phase.

One of the milestones of the project is the creation of an Open Information Hub, a collaborative web based Knowledge Base that will host a catalogue of open tools/services, methodologies, best practices from various disciplines or settings, success stories, reports. More specifically, the OpenUp Info Hub will include:

- a catalogue of open peer review methodologies, initiatives, tools and services; facts and recommendations for metrics and indicators targeted to different stakeholders;
- a directory of innovative dissemination and outreach methods accompanied by good practice guidelines;
- a blog open to the community to host experiences and opinions on any of the OpenUP related aspects;
- a section with user guides, recommendations and FAQs for different categories of stakeholders (young researchers, publishers, funders, policy makers, etc.).

The beta version of the OpenUp Info Hub is presently online and can be visited at: <https://www.openuphub.eu/>.

2.1 Landscaping

This phase determined traditional and ground-breaking mechanisms, processes and tools for peer-review, dissemination, and measuring impact of all types of research results. Using a variety of tools the OpenUP team has scanned the current landscape of traditional and innovative methods, tools and practices across disciplinary, thematic, regional, gender and age borders.

One of the main tools supporting the landscape scans has been a survey conducted between 20 January and 23 February 2017. Following the principles of the project, the questionnaire and the complete data sets of the survey are available at Zenodo, one of the best sites supporting Open Science (<http://doi.org/10.5281/zenodo.556157>).

Peer review: This landscape scan has liaised with similar initiatives (e.g., OpenAIRE's current task on Open Peer Review Systems, which is performing a similar landscaping study, and publishers like F1000 or Frontiers who have advanced ICT enabled peer review systems) and has recorded the processes.

Dissemination: This landscape scan has covered publication approaches as offered by traditional media (e.g. article in newspaper), industrial media (e.g. report as part of a weekly research related magazine) as well as social media (e.g. tweet). OpenUP has also examined and interviewed selected FP7 or H2020 projects to see how they use such dissemination approaches and the impact they gain.

Impact and assessment: This landscape scan has recorded existing and emerging indicators and how they are used in different settings or applications. The main tools here have been surveys and interviews to see what secondary impact indicators (e.g., job growth, societal impact) are important in which setting, and how they can possibly be measured.

Results of the landscaping activity for the three aspects of the research activities (peer review, dissemination and assessment) are available as project deliverables on the OpenUp project web site (<http://openup-h2020.eu/project-materials/project-deliverables>). Content from the deliverables has also been reworked and reformatted to be

included in the OpenUP Info Hub (<https://www.openuphub.eu/>), which provides also an initial set of tools to help implementing the Open Science paradigm.

2.2 Analysing – Framework Phase 1

Based on the landscaping results, OpenUP is presently completing desk analysis to come up with an initial framework for each of the three OpenUP pillars. Specifically, it will produce an interim framework document that will:

- catalogue requirements from different stakeholders
- break down processes to identify commonalities and gaps
- define the qualitative and technical criteria to classify the processes
- define the interrelations among the three pillars and place them within the research workflow.

2.3 Assessing and Validating

During this phase OpenUP will carry out a series of activities to test and validate the proposed innovative mechanisms and indicators against the requirements and needs of key stakeholders (e.g. researchers, funders, innovators, general public). The aim is to deliver first insights into the applicability and practicability of the proposed methods in specific settings and communities, as well as reflect on their effects on the stakeholders involved and on the scientific workflows.

Based on the initial findings, OpenUP is engaged in the rolling-out of seven pilots related to the three pillars, spanning several research communities and initiatives from the life sciences, social sciences, energy, arts and humanity disciplines. The selected communities are: the European Machine Vision Association (EMVA), the eHealth 2018 Student competition, the Human Mortality Database (HMD), DARIAH, Coursera community, the Smarter Together project, and the Berlin Institute of Health. Presently, OpenUP is consulting with the communities to define and refine the implementation and logistics of the pilots to ensure that they reflect the hitherto defined/identified roles, processes, challenges, opportunities as well as identify key questions that may need further investigation.

2.4 Policy Reviewing

The question of how the research findings are (and should be) linked to policy is of direct relevance to OpenUP. Linkages between research and policy may well vary among the three key project pillars, disciplines, research communities and between member States, depending on their overall structuring. It is therefore important to map and analyse the national contexts and existing policies in order to understand areas where the project's findings and recommendations could support evidence-based Research and Innovation policy. OpenUP is presently carrying out several activities to gather and analyse policy data and produce summary reports. In addition to desk research and analysis of available literature, also field research is being carried out, through interviews with policymakers

and survey of key stakeholders in selected countries from the EU-15, EU-13 and Associated Countries (8 countries in total).

2.5 Synthesizing – Framework Phase 2

The last phase of OpenUP will produce a set of practical policy recommendations for EU, national and institutional policymakers for supporting the transition to appropriate and timely measures of quality assurance related to peer review, innovative dissemination of the and their impact measurement. Based on the previous phases, OpenUP will gather all findings (individual frameworks related to the OpenUP pillars, consultations, feedback from validation activities and use cases, policy reviews), will evaluate possible collaborative initiatives between key stakeholders, including researchers, peer reviewers, publishers and policymakers when using the developed approaches and tools to support evidence-informed research and innovation policy. This will be accomplished by: (a) performing a SWOT analysis to propose optimal ways and good practices for implementing the policy in the different European settings and research communities; (b) validating results in focus groups.

3 Work Plan

The project is organized into seven work packages, with the usual structure of the European projects. The relationships among the work packages are shown in Fig. 2.

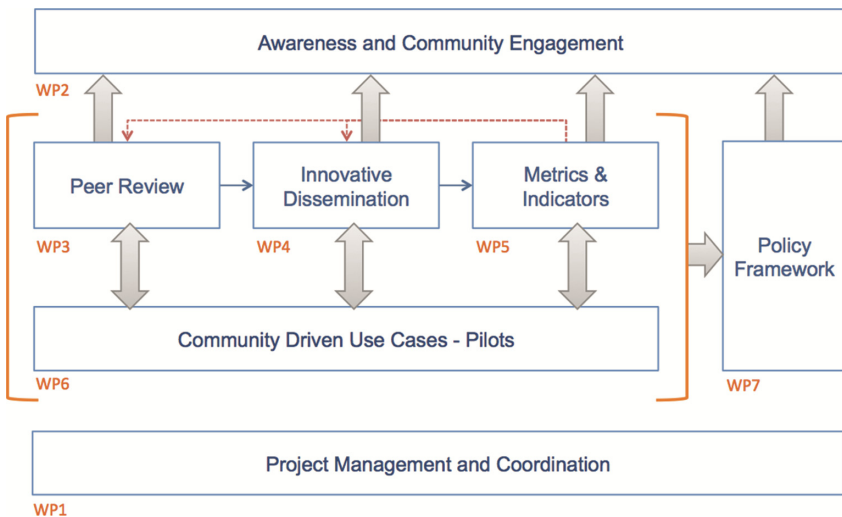


Fig. 2. OpenUP work package relations

WP1 – Management and Coordination is dedicated to the management, coordination and monitoring of the project and enable the efficient progress of its work meeting the contractual obligations and the quality expectancies of the consortium. It also addresses the project’s data management plan and implementation.

WP2 – Outreach and exploitation covers a diverse set of activities that relate to raising awareness about the project in domains of interest and building the instruments for the uptake of the results (framework, pilots, Open Information Hub, recommendations). It also investigates the sustainability model for the long-term operation of the OpenUP communication platform and Information Hub.

WP3 – Peer review framework produces a framework for open peer review on all research artefacts, facilitating a clear definition of the roles and processes, identifying benefits, challenges and opportunities to select questions that need further investigation.

WP4 – Innovative dissemination framework investigates innovative ways of disseminating research outputs beyond traditional academic dissemination in different disciplines, identifying and sharing good practices. The work comes up with practical guidelines on how to create a successful research dissemination strategy beyond traditional academic dissemination.

WP5 – Impact indicators framework generates a validated taxonomy of channels of scientific knowledge dissemination and transfer channels and suggests indicators enabling assessing impact and quality of the underlying research.

WP6 – Community driven use cases and pilots actively engages research communities to validate the frameworks through a set of pilots, eliciting requirements and exploring viable solutions for implementing technical and processual solutions, and getting concrete insights for future research.

WP7 – Policy analysis, recommendations and guidelines is responsible for turning all OpenUP results into practical guidelines and policy recommendations for EU/national/institutional policy makers.

4 The Project

Open UP started in June 2016 and will end in December 2018 (30 months). The nine partners are listed in the Table below. The total cost of the project is about 2.225.000 Euro, with an EU contribution of about 1.950.000 Euro

(see http://cordis.europa.eu/project/rcn/203537_en.html).

All the details of the project can be found at the project web site (<http://openup-h2020.eu/>). As stated above, the results, recommendations and tools developed by the project can also be found at the OpenUp Information Hub (<https://www.openuphub.eu/>)

No.	Participant full organization name	Short name	Country
1	Public Policy and Management Institute (Coordinator)	PPMI	LT
2	Georg-August-Universitaet Stiftung Oeffentlichen Rechts	UGOE	DE
3	National and Kapodistrian University of Athens	UoA	EL
4	Universiteit van Amsterdam	UvA	NL
5	Graz Kompetenzzentrum fur Wissensbasierte Anwendungen und Systeme Forschungs- und Entwicklungs GMBH	KNOW	AT
6	Austrian Institute of Technology	AIT	AT
7	Institut für Forschungsinformation und Qualitätssicherung	IFQ	DE
8	Frontiers Media SA	Frontiers	CH
9	Consiglio Nazionale delle Ricerche	CNR	IT

References

1. Aksnes, D.W., Schneider, J.W., Gunnarsson, M.: Ranking national research systems by citation Indicators. A comparative analysis using whole and fractionalised counting methods. *J. Informetrics* **6**, 36–43 (2012)
2. Aleksic, J., Alexa, A., Attwood, T.K., et al.: An Open Science Peer Review Oath [v2; ref status: indexed, <http://f1000r.es/4wf>, 9 January 2015] *F1000Research*, **3**, 271 (2014). <https://doi.org/10.12688/f1000research.5686.2>
3. Assante, M., Candela, L., Castelli, D., Manghi, P., Pagano, P.: Science 2.0 repositories: time for a change in scholarly communication. *D-Lib Mag.* **21**(1/2) (2015). <https://doi.org/10.1045/january2015-assante>
4. Costas, R., Zahedi, Z., Wouters, P.: Do “altmetrics” correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective? *J. Assoc. Inf. Sci. Technol.* **66**(10), 2003–2019 (2014)
5. Craig, I.D., Plume, A.M., McVeigh, M.E., Pringle, J., Amin, M.: Do open access articles have greater citation impact?: a critical review of the literature. *J. Informetrics* **1**(3), 239–248 (2007)
6. Dinsmore, A., Dolby, K.: Alternative perspectives on impact: The potential of ALMs and altmetrics to inform funders about research impact. *PLoS Biol.* **12** (2014)
7. Egghe, L., Rousseau, R., van Hooydonk, G.: Methods for accrediting publications to authors or countries: Consequences for evaluation studies. *J. Am. Soc. Inf. Sci.* **51**(2), 145–157 (2000)
8. Gauffriau, M., Larsen, P.O.: Counting methods are decisive for rankings based on publication and citation studies. *Scientometrics* **64**(1), 85–93 (2005)
9. Gunn, W.: Social signals reflect academic impact: what it means when a scholar adds a paper to mendeley. *Inf. Stand. Q.* **25**(2), 1–8 (2013). ISSN 1041-0031
10. Guthrie, S., Guérin, B., Wu, H., Sharif I., Wooding, S.: Alternatives to Peer Review in Research Project Funding, RAND report 2013 update. Rand Europe, April 2013
11. Haustein, S., Sugimoto, C.R., Larivière, V.: Social media in scholarly communication. *Aslib J. Inf. Manage.* **67**(3) (2015)

12. Hicks, D., Wouters, P.: The leiden manifesto for research metrics. *Nature* **520**(7548), 429–431 (2015)
13. Langfeldt, L.: The policy challenges of peer review: managing bias, conflict of interests and interdisciplinary assessments. *Res. Eval.* **15**(1), 31–41 (2006). <https://doi.org/10.3152/147154406781776039>
14. Liang, X., Su, L.Y.F., Yeo, S.K., Scheufele, D., Brossard, D., Xenos, M., Corley, E.: Building buzz: (Scientists) communicating science in new media environments. *J. Mass Commun. Q.* **91**(4), 1–20 (2013). <https://doi.org/10.1177/1077699014550092>
15. OpenAIRE: OpenAIRE Open Peer Review Tenders: Selected Projects, Newsletter, 16 September 2015. <https://www.openaire.eu/openaire-open-peer-review-tenders>
16. Peroni, S., Dutton, A., Gray, T., Shotton, D.: Setting our bibliographic references free: towards open citation data. *J. Documentation* **71**(2), 253–277 (2015)
17. Ponte, D., Simon, J.: Scholarly communication 2.0: Exploring researchers' opinions on web 2.0 for scientific knowledge creation, evaluation and dissemination. *Serials Rev.* **37**(3), 149–156 (2011). <https://doi.org/10.1080/00987913.2011.10765376>
18. Pöschl, U.: Multi-stage open peer review: scientific evaluation integrating the strengths of traditional peer review with the virtues of transparency and self-regulation. *Front. Comput. Neurosci.* **6**(33) (2012). <https://doi.org/10.3389/fncom.2012.00033>
19. Procter, R., Williams, R., Stewart, J.: If you Build it, Will They Come? A Research Information Network report, July 2010. http://www.rin.ac.uk/system/files/attachments/web_2.0_screen.pdf
20. Roemer, R.C., Borchardt, R.: From bibliometrics to altmetrics. *Coll. Res. Libr. News* **73**(10), 596–600 (2012)
21. Sotudeh, H., Ghasempour, Z., Yaghtin, M.: The citation advantage of author-pays model: the case of Springer and Elsevier OA journals. *Scientometrics* **104**, 581–608 (2015)
22. Su, L.Y.-F., Akin, H., Brossard, D., Scheufele, D.A., Xenos, M.A.: Science news consumption patterns and their implications for public understanding of science. *J. Mass Commun. Q.* (2015). <https://doi.org/10.1177/1077699015586415>
23. Waltman, L., Van Eck, N.J.: The inconsistency of the h-index. *J. Am. Soc. Inform. Sci. Technol.* **63**(2), 406–415 (2012)

Who Is the Data Curator? Defining a Vocabulary

Anna Maria Tammaro¹ and Vittore Casarosa²(✉)

¹ University of Parma, Parma, Italy

² ISTI-CNR, Pisa, Italy

casarosa@isti.cnr.it

Abstract. In 2016, the IFLA Section Library Theory and Research has (partially) funded the research project “Data curator who is s/he?” to clarify the profile of data curator. The main goal of the project was to define characteristics of roles and responsibilities of data curators in the international and interdisciplinary contexts. The research questions of the Project were:

R1: How is data curation defined by practitioners/professional working in the field?; R2: What terms are used to describe the roles for professionals in data curation area?; R3: What are primary roles and responsibilities of data curators?; R4: What are educational qualifications and competencies required of data curators?

In this paper we present briefly some of the results related to research questions R1 and R2, namely what terms are used to describe the roles for professionals in data curation area.

Keywords: Data curator · Data curation · Vocabulary

1 Introduction

Data curation has been defined by the University of Illinois as “the active and ongoing management of data through its life cycle of interest and usefulness to scholarship, science, and education. Data curation activities enable data discovery and retrieval, maintain its quality, add value, and provide for reuse over time, and this new field includes authentication, archiving, management, preservation, retrieval, and representation.”

After the introduction of the term “curation” by the DCC centre, an interdisciplinary professional community has investigated the role and responsibilities of a new or renewed profile called “data curator”. Data curators are specialists who play an important role, whose definition is still in progress and not agreed. The unique feature of the new profile is that it is considered a necessity in all countries; however, competencies, roles and qualifications highlight different backgrounds. In addition, the profile is evolving from the reflective practice of professionals, and there is not a shared theory for the role. The activities of the new profile are also unclear and there is no professional association that has analysed and prepared a competence list (see selected bibliography in the References Section).

In 2016, the IFLA Section Library Theory and Research has (partially) funded the research project “Data curator who is s/he?” to clarify the profile of data curator.

The main goal of the project was to define characteristics of roles and responsibilities of data curators in the international and interdisciplinary contexts.

The methodology used in the IFLA Project has been based on a mixed-methodology, with content analysis of job announcements for data curators and for librarians to be involved in Research Data Management activities on one side, and semi-structured interviews with professionals working as data librarians, data curators, or research data managers on the other.

The research questions of the Project were:

- R1: How is data curation defined by practitioners/professional working in the field?
- R2: What terms are used to describe the roles for professionals in data curation area?
- R3: What are primary roles and responsibilities of data curators?
- R4: What are educational qualifications and competencies required of data curators?

The project, funded by the IFLA, was articulated in three main phases:

- Phase I – Literature review and vocabulary.
- Phase II – Content analysis of job announcements (appeared in the time frame 2015–2017) offering positions with data curation responsibilities in libraries, archives, and research centers. The job offers were collected from:
 - Job postings from the American Library Association (<http://joblist.ala.org/>)
 - The community driven site of Code4lib (<http://jobs.code4lib.org/jobs/data-curation/>)
 - The IASSIST Jobs Repository (<http://www.iassistdata.org/resources/jobs/all>)
- Phase III – Document analysis of interviews with data curators and questionnaires distributed to data curators. All interviews were conducted in 2017 and in several geographic areas.

In this paper we present briefly some of the results related to the research questions R1 and R2, namely what terms are used to describe the roles for professionals in the data curation area. Although it is a partial result, it seemed interesting to us to highlight the methodology of building the ontology, and how the ontology has allowed us to analyse both conceptual differences between communities and the growth of profile understanding over time between 2015 and 2017. This was an unexpected result. To monitor the development of the profile and, above all, to collaborate with data scientist and computer community scientists to arrive at an agreed terminology, the ontology remains open. The project results related to interviews and job description are going to be published in the final report.

2 Term Extraction from Corpora Related to Digital Curation/Digital Curator

One of the main intent of the project was to identify a set of terms (a vocabulary) and possibly an ontology related to Digital Curation, by analysing relevant textual data in the field. Six different corpora were collected, identified with the following nicknames.

- Bibliography Old. Text extracted from abstracts and keywords of papers related to Digital Curation and published up to 2015.
- Bibliography New. Text extracted from abstracts and keywords of papers related to Digital Curation and published in 2016 and 2017.
- Positions/Job offers. Text extracted from job offers and positions, searching for “digital curators”, mostly from academia.
- Questionnaire. Text extracted from a set of questionnaires distributed to professionals already working as data librarians, data curators, or research data managers.
- Interviews. Text extracted from the transcript of interviews with selected respondents of the questionnaires.
- Edison project. Text extracted from deliverables of the Edison project, a European project aiming at defining the skills and the roles of the new profession of Data Scientist.

The system used to extract relevant terms from the corpora, more generally called “key phrases” is the Keyphrase Digger (KD, see <http://dh.fbk.eu/technologies/kd>), developed at the Fondazione Bruno Kessler (FBK, see <http://ict.fbk.eu/>). KD scans a given corpus, and computes the “scores” of candidate key phrases, based on term frequency measures and linguistic syntactic information (Part of Speech patterns). It then return the key phrases in descending order of their score. The Keyphrase Digger has three main parameters:

- n, the number of key phrases to be returned, which was set to 50 for all corpora except the Interviews. In this case, given the highly unstructured and colloquial text, n was set to 80 in order to try and capture more relevant key phrases.
- p, which gives a boost to more specific key-concepts (ie. multi-token expressions). Depending on the value of p there will be more or less multi-token expressions in the result. It can have values NO, WEAK, MEDIUM, STRONG.
- m, which indicated the maximum number of words that can be used in the multi-token expressions.

In order to try and extract the maximum amount of information, KD was run on each corpus several times, with different values of the p and m parameters. A first series of five runs was done, with p = WEAK and m (the maximum number of words in a key phrase) going from 1 to 5. Then the results of the five runs were merged into a single list eliminating duplicates and terms clearly not related to Digital Curation (based on subjective judgement). It has to be noted here that KD does not have any “semantic” knowledge, and the key phrases identified are based only on frequencies and syntactic information. Therefore the returned list may contain “key phrases” not related at all with Digital Curation, even if the system has a stop word list to eliminate the most common words.

The same process of five runs was applied to the same corpus with p = STRONG, obtaining a second list of relevant key phrases. The two lists were finally merged into a single one, eliminating duplicate terms and obtaining the set of terms related to Digital Curation for that corpus.

After having extracted the set of key phrases from the six different corpora, it was apparent that there was a minimal overlap between any two pairs of corpora. To make this observation more precise, we generated the intersection of the 15 possible pairs of corpora, obtaining the key phrases in common between any two sets. In Appendix A there is the table with all the overlapping terms. The results confirmed the initial observation, as the “intersection sets” go from a minimum of 3 elements for the pair Interviews/Edison to a maximum of 18 elements for the pair Bibliography New/Bibliography Old.

This result may not be surprising, if we consider that the texts in the corpora are coming from different communities, and if we assume that each community may have its own terminology. What remains to be understood (possibly in a continuation of the project) is whether the differences are just a matter of “terminology”, i.e. different communities use different terms to indicate (more or less) the same set of concepts, or is rather a difference in the set of concepts related to Digital Curation, assuming that the “relevance” of a concept (and therefore its appearance in the final results of KD) is actually depending on the community using it.

3 Data Curation Concepts

To start trying to understand if and to what extent the overlapping in terminology between the different corpora is just a matter of terminology or is rather a matter of concepts, we have matched some of the common terms extracted from the corpora with the definitions found in Wikidata. We have also started building a table of “related terms” in order to arrive to a more complete and articulated taxonomy in the field of Data Curation. The initial preliminary results are shown below.

- Data curation (Q15088675)
work performed to ensure meaningful and enduring access to data
- Data management (Q1149776)
all disciplines related to managing data as a valuable resource
- Digital curation (Q5276060)
selection, preservation, maintenance, collection and archiving of digital assets
- Digital Preservation (Q632897)
formal endeavor to ensure that digital information of continuing value remains accessible and usable
- Preservation (Q1479406)
maintenance of objects as closely as possible to their original condition, also called conservation
- Research Data (Q15809982)
collection of facts produced through systematic inquiry
- Research Data Management RDM (Q30089794)
activities around the life cycle of research-related data

Term	Definition	Related term	Code
Research Data Management (RDM)	Activities around the life cycle of research-related data	Research data: collection of facts produced through systematic inquiry (Q15809982)	(Q30089794)
Data curation	Work performed to ensure meaningful and enduring access to data	Digital curation: selection, preservation, maintenance, collection and archiving of digital assets (Q5276060)	(Q15088675)
Data management	All disciplines related to managing data as a valuable resource	Data management plan (Q17085509)	(Q1149776)
Digital preservation	Formal endeavor to ensure that digital information of continuing value remains accessible and usable	Preservation: maintenance of objects as closely as possible to their original condition also called conservation (Q1479406)	(Q632897)
Data science	Interdisciplinary field about processes and systems to extract knowledge or insights from data	Data scientist: a person studying and working with data (Q29169143)	(Q2374463)

Appendix A

	Bibliography old	Bibliography new	Positions/Job offers	Questionnaire	Interviews	Edison
Bibliogr. Old (60 key phrases)		curation//data curation//data curation education//data curation profiles//data management//data quality//digital curation//digital preservation// escience professionals//large scale information management problems// metadata//plurality of curation roles// preservation// repositories// research data// research data curation//research data management	curation//data curation//data management// digital curation// digital preservation// preservation// research data// research data curation//research data management	curation//data curation//data management//data quality//data sharing//data stewardship// digital curation// digital preservation// preservation// repositories// research data// research data management	area of data// curation//data curation//data management//data management plan// digital curation// metadata//research data//research data management	curation//data curation//data management//data quality

(continued)

(continued)

	Bibliography old	Bibliography new	Positions/Job offers	Questionnaire	Interviews	Edison
Bibliogr. New (69 key phrases)	curation//data curation//data curation education//data curation profiles//data management//data quality//digital curation//digital preservation// escience professionals//large scale information management problems// metadata//plurality of curation roles// preservation// repositories// research data// research data curation//research data management		curation//data curation//data management//data management planning//digital curation//digital preservation// preservation// research data// research data curation//research data management	curation//data curation//data management//data management planning//data quality//digital curation//digital preservation// preservation// information management// preservation// repositories// research data// research data management// research data services	curation// curators//data curation//data management// digital curation// metadata//research data//research data management	curation//data curation//data management//data quality//data science
Positions (40 key phrases)	curation//data curation//data management// digital curation// digital preservation// preservation// research data// research data curation//research data management	curation//data curation//data management//data management planning//digital curation//digital preservation// preservation// research data// research data curation//research data management		curation//data curation//data management//data management planning//digital curation//digital preservation// preservation// research data// research data curation//research data management	curation//data curation//data management// digital curation// research data// research data management// scholarly communication	curation//data curation//data management
Question. (56 key phrases)	curation//data curation//data management//data quality//data sharing//data stewardship// digital curation// digital preservation// preservation// repositories// research data// research data management	curation//data curation//data management//data management planning//data quality//digital curation//digital preservation// information management// preservation// repositories// research data// research data management// research data services	curation//data curation//data management//data management planning//digital curation//digital preservation// preservation// research data// research data management		curation//data curation//data management// digital curation// research data// research data management	curation//data curation//data management//data quality
Interviews (34 key phrases)	area of data/ curation//data curation//data management//data management plan// digital curation// metadata//research data//research data management	curation// curators//data curation//data management// digital curation// metadata//research data//research data management	curation//data curation//data management// digital curation// research data// research data management// scholarly communication	curation//data curation//data management// digital curation// research data// research data management		curation//data curation//data management
Edison p(25 key phrases)	curation//data curation//data management//data quality	curation//data curation//data management//data quality//data science	curation//data curation//data management	curation//data curation//data management//data quality	curation//data curation//data management	

References

1. Akers, K.G., Sferdean, F.C., Nicholls, N.H., Green, J.A.: Building support for research data management: biographies of eight research universities. *Int. J. Digit. Curation* **9**(2), 171–191 (2014). <https://doi.org/10.2218/ijdc.v9i2.327>
2. Bailey, C.W.: *Research Data Curation Bibliography*. Digital Scholarship, Houston (2017). <http://digital-scholarship.org/rdcb/rdcb.htm>
3. Beagrie, N.: Digital curation for science, digital libraries, and individuals. *Int. J. Digit. Curation* **1**(1), 3–16 (2008). <https://doi.org/10.2218/ijdc.v1i1.2>
4. Heidorn, P.B.: The emerging role of libraries in data curation and e-science. *J. Libr. Adm.* **51** (7–8), 662–672 (2011)
5. Kim, J., Warga, E., Moen, W.E.: Digital curation in the academic library job market. *Proc. Am. Soc. Inf. Sci. Technol.* **49**(1), 1–4 (2012). <https://doi.org/10.1002/meet.14504901267>
6. Kim, J., Warga, E., Moen, W.E.: Competencies required for digital curation: an analysis of job advertisements. *Int. J. Digi. Curation* **8**(1), 66–83 (2013). <https://doi.org/10.2218/ijdc.v8i1.242>
7. Lee, C.A., Tibbo, H.R., Schaefer, J.C.: Defining what digital curators do and what they need to know: the DigCCURR project. In: *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 49–50. ACM, New York (2007). <https://doi.org/10.1145/1255175.1255183>
8. Molloy, L., Gow, A., Konstantelos, L.: The DigCurV curriculum framework for digital curation in the cultural heritage sector. *Int. J. Digit. Curation* **9**(1), 231–241 (2014)
9. Palmer, C.L., Thompson, C.A., Baker, K.S., Senseney, M.: Meeting data workforce needs: indicators based on recent data curation placements. In: *iConference 2014 Proceedings* (2014). <http://hdl.handle.net/2142/47308>
10. Poole, A.H.: Now is the future now? The urgency of digital curation in the digital humanities. *Digit. Humanit. Q.* **7**(2). <http://www.digitalhumanities.org/dhq/vol/7/2/000163/000163.html>
11. Pryor, G.: *Why manage research data?* In: Pryor, G. (ed.) *Managing Research Data*. Facet Publishing, London (2012)
12. Tamaro, A.M., Matusiak, K., Sposito, F.A., Pervan, A., Casarosa, V.: Understanding roles and responsibilities of data curators: an international perspective. *Libellarium* **9**(2), 39–47 (2016)
13. Tenopir, C., Sandusky, R.J., Allard, S., Birch, B.: Research data management services in academic research libraries and perceptions of librarians. *Libr. Inf. Sci. Res.* **36**(2), 84–90 (2014)
14. Walters, T., Skinner, K.: New roles for new times: digital curation for preservation. *Association of Research Libraries, Washington, DC* (2011). http://www.arl.org/storage/documents/publications/nrnt_digital_curation17mar11.pdf
15. Witt, M.: Institutional repositories and research data curation in a distributed environment. *Libr. Trends* **57**(2), 191–201 (2008). <https://doi.org/10.1353/lib.0.0029>

Author Index

- Adorni, Giovanni 114
Agosti, Maristella 30, 42
Allavena, Davide 15
Allegrezza, Stefano 209
Antolli, Elisa 180
- Baecchi, Claudio 163
Balducci, Fabrizio 120
Baraldi, Lorenzo 169
Barbero, Giliola 201
Barbuti, Nicola 139
Bardi, Alessia 240
Bartalesi, Valentina 23
Basaldella, Marco 93, 180
Benedetti, Filippo 23
Bertini, Marco 163
Bevilacqua, Giorgio 15
Bolelli, Federico 151
Borghi, Guido 151
Bressan, Federica 101
- Caldarola, Tommaso 139
Casarosa, Vittore 240, 249
Cerullo, Luigi 219
Cornia, Marcella 169
Cucchiara, Rita 169
Cullhed, Per 81
Cuna, Andrea 3
- Daud Awan, Muhammad 127
Del Bimbo, Alberto 163
Dell'Orletta, Felice 114
Di Matteo, Nicola R. 209
Di Nunzio, Giorgio Maria 30, 42
- Esposito, Floriana 69
- Ferilli, Stefano 57, 69, 139
Ferracani, Andrea 163
Ferro, Nicola 30, 42
- Grana, Costantino 151
- Hast, Anders 81
- Khan, Muzammil 127
Koceva, Frosina 114
- Lana, Maurizio 191
Leman, Marc 101
- Maistro, Maria 42
Manghi, Paolo 240
Marchesin, Stefano 42
Marlazzi, Petra 225
Meghini, Carlo 23
Metilli, Daniele 23
Morando, Federico 15
- Orio, Nicola 42
- Parolo, Lisa 225
Pazienza, Andrea 57
Pini, Stefano 169
Ponchia, Chiara 42
- Saba, Cosetta 225
Schiavone, Luisa 15
Seidenari, Lorenzo 163
Serra, Giuseppe 93, 180
Silvello, Gianmaria 30, 42
Six, Joren 101
- Tambassi, Timothy 191
Tammaro, Anna Maria 249
Tasso, Carlo 93, 180
Tessarolo, Luigi 201
Torre, Ilaria 114
- Ur Rahman, Arif 127
Uricchio, Tiberio 163
- Vats, Ekta 81
Venturi, Giulia 114
Vitacolonna, Nicola 225