Managing the Knowledge Creation Process of Large-Scale Evaluation Campaigns

Marco Dussin and Nicola Ferro

University of Padua, Italy
{dussinma,ferro}@dei.unipd.it

Abstract. This paper discusses the evolution of large-scale evaluation campaigns and the corresponding evaluation infrastructures needed to carry them out. We present the next challenges for these initiatives and show how digital library systems can play a relevant role in supporting the research conducted in these fora by acting as virtual research environments.

1 Introduction

Large-scale evaluation initiatives provide a significant contribution to the building of strong research communities, advancement in research and state-of-the-art, and industrial innovation in a given domain. Relevant and long-lived examples from the Information Retrieval (IR) field are the Text REtrieval Conference (TREC)[1] in the United States, the Cross-Language Evaluation Forum (CLEF)[2] in Europe, and the NII-NACSIS Test Collection for IR Systems (NTCIR)[3] in Japan and Asia. Moreover, new initiatives are growing to support emerging communities and address specific issues, such as the Forum for Information Retrieval and Evaluation (FIRE)[4] in India.

These initiatives impact not only the IR field itself but also related fields which adopt and apply results from it, such as the Digital Library (DL) one. Indeed, the information access and extraction components of a DL system, which index, search and retrieve documents in response to a user's query, rely on methods and techniques taken from the IR field. In this context, large-scale evaluation campaigns provide qualitative and quantitative evidence over the years as to which methods give the best results in certain key areas, such as indexing techniques, relevance feedback, multilingual querying, and results merging, and contribute to the overall problem of evaluating a DL system [14].

This paper presents a perspective on the evolution of large-scale evaluation campaigns and their infrastructures and the challenges that they will have to face in the future. The discussion will provide the basis to show how these emerging challenges call for an appropriate consideration and management of the

[1] `http://trec.nist.gov/`
[2] `http://www.clef-campaign.org/`
[3] `http://research.nii.ac.jp/ntcir/`
[4] `http://www.isical.ac.in/~clia/`

M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 63–74, 2009.

knowledge creation process involved by these initiatives, and how DL systems can play an important role in the evolution of large-scale evaluation campaigns and their infrastructures by acting as virtual research environments.

The paper is organized as follows: Section 2 summarizes the evolution of large-scale evaluation campaigns, the challenges for their future and our vision of the extension to the current evaluation methodology to address these challenges; Sections 3 and 4 discuss the DIKW hierarchy as a means of modeling the knowledge creation process of an evaluation campaign; Section 5 presents the DIRECT digital library system to show how the previously introduced concepts can be applied; finally, Section 6 draws some conlusions.

2 Evolution of Large-Scale Evaluation Campaigns and Infrastructures

Large-scale evaluation campaigns have been a driver of research and innovation in IR since the early 90s, when TREC was launched [15]. They have been relying mainly on the traditional Cranfield methodology [6], which focuses on creating comparable experiments and evaluating their performance. During their life span, large-scale evaluation campaigns have produced a great amount of research not only on specific IR issues – such has indexing schemes, weighting functions, retrieval models, and so on – but also on improving the evaluation methodology itself, for example, with respect to the effective and efficient creation of reliable and re-usable test collections, the proposal and study of appropriate metrics for assessing a task, or the application of suitable statistical techniques to validate and compare the results.

As part of recent efforts to shape the future of large-scale evaluation campaigns [3,12], more attention has been paid to evaluation infrastructures, meant as the information management systems that have to take care of the different steps and outcomes of an evaluation campaign. The need for appropriate evaluation infrastructures which allows for better management and exploitation of the experimental results has been highlighted also by different organizations, such the European Commission in the i2010 Digital Library Initiative [10], the US National Scientific Board [16], and the Australian Working Group on Data for Science [19].

In this context, we have proposed an extension to the traditional evaluation methodology in order to explicitly take into consideration and model the valuable scientific data produced during an evaluation campaign [2,5], the creation of which is often expensive and not easily reproducible. Indeed, researchers not only benefit from having comparable experiments and a reliable assessment of their performances, but they also take advantage of the possibility of having an integrated vision of the scientific data produced, together with their analyses and interpretations, as well as benefiting from the possibility of keeping, re-using, preserving, and curating them. Moreover, the way in which experimental results are managed, made accessible, exchanged, visualized, interpreted, enriched and referenced is therefore an integral part of the process of knowledge transfer and sharing towards relevant application communities, such as the DL one, which needs to properly understand these experimental results in order to create and assess their own systems.

Therefore, we have undertaken the design of an evaluation infrastructure for large-scale evaluation campaigns and we have chosen to rely on DL systems in order to develop it, since they offer content management, access, curation, and enrichment functionalities. The outcome is a DL system, called Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)[5], which manages the scientific data produced during a large-scale evaluation campaign, as well as supports the archiving, access, citation, dissemination, and sharing of the experimental results [7,8,9]. DIRECT has been used, developed and tested in the course of the annual CLEF campaign since 2005.

2.1 Upcoming Challenges for Large-Scale Evaluation Campaigns

Since large-scale evaluation campaign began, the associated technologies, services and users of information access systems have been in continual evolution, with many new factors and trends influencing the field. For example, the growth of the Internet has been exponential with respect to the number of users and languages used regularly for global information dissemination. With the advance of broadband access and the evolution of both wired and wireless connection modes, users are now not only information consumers, but also information producers: creating their own content and augmenting existing material through annotations (e.g. adding tags and comments) and cross-referencing (e.g. adding links) within a dynamic and collaborative information space. The expectations and habits of users are constantly changing, together with the ways in which they interact with content and services, often creating new and original ways of exploiting them. Moreover, users need to be able to co-operate and communicate in a way that crosses language boundaries and goes beyond simple translation from one language to another. Indeed, language barriers are no more perceived simply as an "obstacle" to retrieval of relevant information resources, they also represent a challenge for the whole communication process (i.e. information access and exchange). This constantly evolving scenario poses new challenges to the research community which must react to these new trends and emerging needs.

From a glance at Figure 1, it can be noted that large-scale evaluation campaigns initially assumed a user model reflecting a simple information seeking behavior: the retrieval of a list of relevant items in response to a single query that could then be used for further consultation in various languages and media types. This simple scenario of user interaction has allowed researchers to focus their attention on studying core technical issues for information access systems and associated components. If we are to continue advancing the state-of-the-art in information access technologies, we need to understand a new breed of users, performing different kinds of tasks within varying domains, often acting within communities to find and produce information not only for themselves, but also

[5] http://direct.dei.unipd.it/

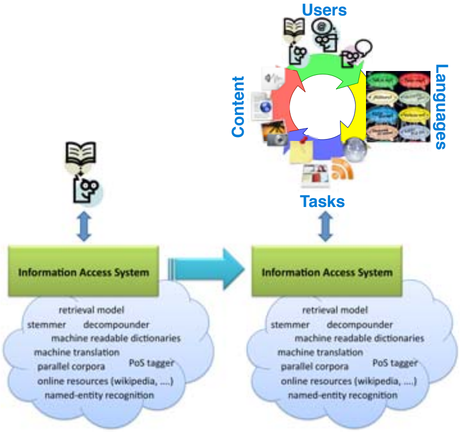

Fig. 1. The evolution of information access technologies

to share with other users. To this end, we must study the interaction among four main entities: users, their tasks, languages, and content to help understand how these factors impact on the design and development of information access systems.

2.2 Upcoming Challenges for Large-Scale Evaluation Infrastructures

The future challenges for the evaluation campaigns will require an increased attention for the knowledge process entailed by an evaluation campaign. The complexity of the tasks and the interactions to be studied and evaluated will produce, as usual, valuable scientific data, which will provide the basis for the analyses and need to be properly managed, curated, enriched, and accessed. Nevertheless, to effectively investigate these new domains, not only the scientific data but also the information and knowledge derived from them will need to be appropriately treated and managed, as well as the cooperation, communication, discussion, and exchange of ideas among researchers in the field. As a consequence, we have to further advance the evaluation methodologies in order to support the whole knowledge creation process entailed by a large-scale evaluation campaign and to deal with the increasing complexity of the tasks to be evaluated. This requires the design and development of evaluation infrastructures which offer better support for and facilitate the research activities related to an evaluation campaign.

A first step in this direction, which is also the contribution of the paper, is to approach and study the information space entailed by an evaluation campaign in the light of the Data, Information, Knowledge, Wisdom (DIKW) hierarchy [1,20], used as a model to organize the information resources produced during it. The study contributes to creating awareness about the different levels and increasing complexity of the information resources produced during an evaluation campaign and indicates the relationships among the different actors involved in it, their tasks, and the information resources produced. The outcomes of this study are then applied in the design and development of the DIRECT system in order to validate their usefulness and effectiveness in the context of CLEF, which represents a relevant example of large-scale evaluation campaign with about 100 participating research groups per year.

In the perspective of the upcoming challenges, our final goal is to turn the DIRECT system from a DL for scientific data into a kind of virtual research environment, where the whole process which leads to the creation, maintenance, dissemination, and sharing of the knowledge produced during an evaluation campaign is taken into consideration and fostered. The boundaries between *content producers* – evaluation campaign organizers who provide experimental collections, participants who submit experiments and perform analyses, and so on – and *content consumers* – students, researchers, industries and practicioners who use the experimental data to conduct their own research or business, and to develop their own systems – are lowered by the current technologies: considering that we aim at making DIRECT an active communication vehicle for the communities interested in the experimental evaluation. This can be achieved by extending the DL for scientific data with advanced annotation and collaboration functionalities in order to become not only the place where storing and accessing the experimental results take place, but also an active communication tool for studying, discussing, comparing the evaluation results, where people can enrich the information managed through it with their own annotations, tags, ... and share them in a sort of social evaluation community. Indeed, the annotation of digital content [4,11] which ranges from metadata, tags, bookmarks, to comments and discussion threads, is the ideal means for fostering the active involvement of user communities and is one of the advanced services which the next generation digital libraries aim at offering.

3 The DIKW Hierarchy

The Data, Information, Knowledge, Wisdom (DIKW) hierarchy is a widely recognized model in the information and knowledge literature [1,18,20]. The academic and professional literature supports diversified meanings for each of the four concepts, discussing the number of elements, their relations, and their position in the structure of hierarchy. In particular, [18] summarizes the original articulation of the hierarchy and offers a detailed and close examination of the similarities and differences between the subsequent interpretations, and [13] identifies the good and the bad assumptions made about the components of the hierarchy. The four layers can summarized as follows:

- at the *data layer* there are raw, discrete, objective, basic elements, partial and atomized, which have little meaning by themselves and no significance beyond their existence. Data are defined as symbols that represents properties of objects, events and their environment, are created with facts, can be measured, and can be viewed as the building blocks of the other layers;
- the *information layer* is the result of computations and processing of the data. Information is inferred from data, answers to questions that begin

with *who, what, when* and *how many*. Information comes from the form taken by the data when they are grouped and organized in different ways to create relational connections. Information is data formatted, organized and processed for a purpose, and it is data interpretable and understandable by the recipient;

- the *knowledge layer* is related to the generation of appropriate actions, by using the appropriate collection of information gathered at the previous level of the hierarchy. Knowledge is *know what* and *know that*, articulable into a language, more or less formal, such as words, numbers, expressions and so on, and transmittible to others (also called *explicit knowledge* [17]), or *know how*, not necessarily codifiable or articulable, embedded in individual experience, like beliefs or intuitions, and learned only by experience and communicated only directly (*tacit knowledge* [17]).
- the *wisdom layer* provides interpretation, explanation, and formalization of the content of the previous levels. Wisdom is the faculty to understand how to apply concepts from one domain to new situations or problems, the ability to increase effectiveness, and it adds value by requiring the mental function we call judgement. Wisdom is not one thing: it is the highest level of understanding, and a uniquely human state. The previous levels are related to the past, whereas with wisdom people can strive for the future.

Those four layers can be graphically represented as a continuum linear chain or as the *knowledge pyramid*, where the wisdom is identified as the pinnacle of the hierarchy, and it is possible to see some transitions between each level in both directions [18]. There is a consensus that data, information, and knowledge are to be defined in terms of one another, but less agreement as to the conversion of one into another one. According to [18], moreover, wisdom is a very elusive concept in the literature about DIKW hierarchy, because the a limited discussion of its nature, "and even less discussion of the organizational processes that contribute to the cultivation of wisdom", despite its position at the pinnacle of the hierarchy.

4 Applying the DIKW Hierarchy to Large-Scale Evaluation Campaigns

Our aim is to define a relationship between the elements of the DIKW hierarchy and the knowledge process carried out by the actors involved in an evaluation campaign. Indeed, each step of a campaign and its outcomes can be coupled with specific actors and with one or more elements of the hierarchy. The result is a chain linking each step with a particular information resource, such as *experiments*, *performance measurements*, *papers*, etc., and the actors involved. Note that wisdom "has more to do with human intuition, understanding, interpretation and actions, than with systems" [18], but passing through the chain, each campaign become a spiral staircase connected to the other campaigns, allowing the user to create their own path to move towards wisdom supported by a system able to support and make explicit each step.

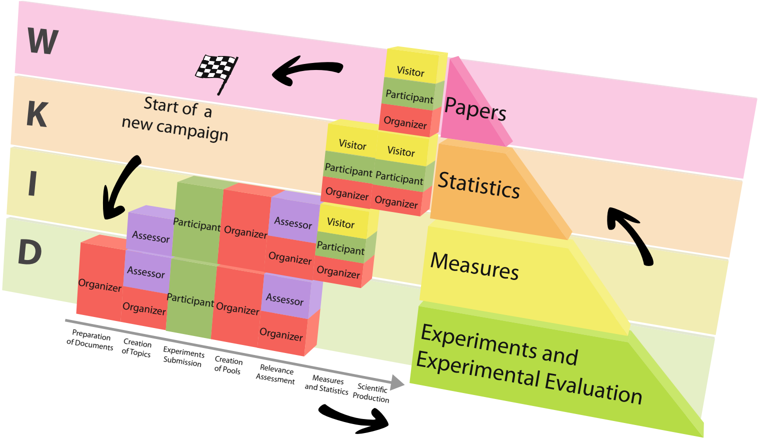

Fig. 2. DIKW knowledge pyramid applied to large-scale evaluation campaigns

Figure 2 frames the different types of information resources, actors, and main steps involved in an evaluation campaign into the pyramid of the DIKW hierarchy.

The left facet of the pyramid is created by a table that summarizes the relationships between the main steps of an evaluation campaign, shown in chronological order on the horizontal axis, the elements of the DIKW hierarchy, shown on the vertical axis, and the main actors involved in an evaluation campaign. For practical reasons, the D, I, K, and W layers are represented as separated, but each step can produce resources that belong to more than one layer.

The right facet summarizes the information resources given at the end of the campaign at each level of the hierarchy: in this way, we can talk about the *experimental collections* and the *experiments* as *data*, since they are raw elements: in fact, an experiment is useless without a relationship with the experimental collection with respect to which the experiment has been conducted. The *performance measurements*, by associating meaning to the data through some kind of relational connection, and being the result of computations and processing on the data, are *information*; the *descriptive statistics* and the *hypothesis tests* are *knowledge* since they are carried by the performance measurements and could be used to make decisions and take further actions about the scientific work. Finally, *wisdom* is provided by *theories, models, algorithms, techniques*, and *observations*, communicated by means of papers, talks, and seminars to formalize and explain the content of the previous levels.

The arrows in Figure 2 explain how each campaign is a step of a cycle where information resources generated in the past are used to allow the user to move towards wisdom as on a spiral staircase. The role of different actors is central to this process since their interactions make it possible to pass from one layer to another.

5 The DIRECT Digital Library System

DIRECT has successfully adopted in the CLEF campaigns since 2005 and has allowed us to:

– CLEF 2005: manage 530 experiments submitted by 30 participants spread over 15 nations and assess more than 160,000 documents in seven different languages, including Bulgarian and Russian which use the Cyrillic alphabet, thanks to the work of 15 assessors;
– CLEF 2006: manage 570 experiments submitted by 75 participants spread over 25 nations and assess more than 200,000 documents in nine different languages, thanks to the work of 40 assessors;
– CLEF 2007: manage 430 experiments submitted by 45 participants spread over 18 nations and assess more than 215,000 documents in seven different languages, thanks to the work of 75 assessors;
– CLEF 2008: manage 490 experiments submitted by 40 participants spread over 20 nations and assess more than 250,000 documents in seven different languages, including Farsi which is written from right to left, thanks to the work of 65 assessors.

In the following, we present the architecture and one example of the functionalities of the DIRECT system.

5.1 Architecture

DIRECT has been designed to be cross-platform and easily deployable to end users; to be as modular as possible, clearly separating the application logic from the interface logic; to be intuitive and capable of providing support for the various user tasks described in the previous section, such as experiment submission, consultation of metrics and plots about experiment performances, relevance assessment, and so on; to support different types of users, i.e. participants, assessors, organizers, and visitors, who need to have access to different kinds of features and capabilities; to support internationalization and localization: the application needs to be able to adapt to the language of the user and their country or culturally dependent data, such as dates and currencies.

Figure 3 shows the architecture of the system. It consists of three layers:

– *data logic*: this deals with the persistence of the different information objects coming from the upper layers. There is a set of "storing managers" dedicated to storing the submitted experiments, the relevance assessments and so on. The Data Access Object (DAO) pattern implements the access mechanism required to work with the underlying data source, acting as an adapter between the upper layers and the data source. Finally, on top of the various DAOs there is the "DIRECT Datastore" which hides the details about the storage management to the upper layers. In this way, the addition of a new DAO is totally transparent for the upper layers.
– *application logic*: this layer deals with the flow of operations within DIRECT. It provides a set of tools capable of managing high-level tasks, such as experiment submission, pool assessment, and statistical analysis of an experiment. For example, the "Performance Measures and Statistical Analyses" tool offers the functionalities needed to conduct a statistical analysis on a set of experiments. In order to ensure comparability and reliability, the tool makes

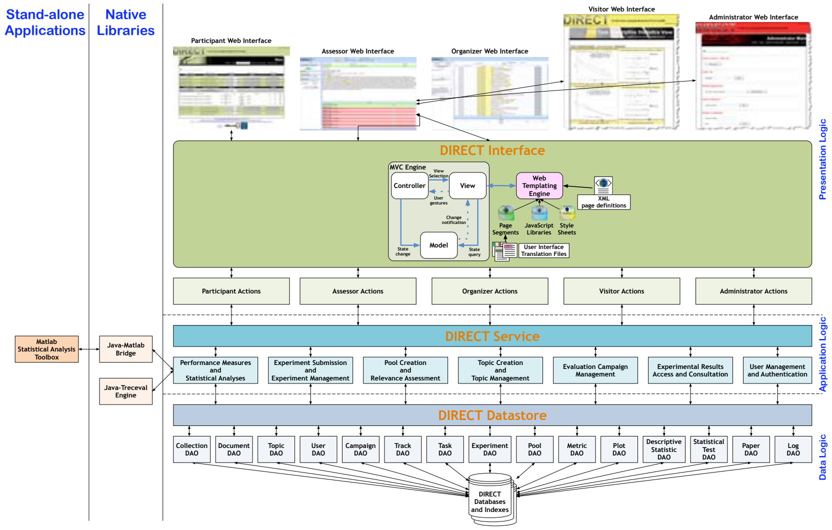

Fig. 3. Architecture of the DIRECT system

uses of well-known and widely used tools to implement the statistical tests, so that everyone can replicate the same test, even if they have no access to the service. In the architecture, the MATLAB Statistics Toolbox has been adopted, since MATLAB is a leader application in the field of numerical analysis which employs state-of-the-art algorithms, but other software could have been used as well. Finally, the "DIRECT Service" provides the interface logic layer with uniform and integrated access to the various tools. As in the case of the "DIRECT Datastore", thanks to the "DIRECT Service" the addition of new tools is transparent for the interface logic layer.

– *interface logic*: this is a Web-based application based on the Model-View-Controller (MVC) approach in order to provide modularity and a clear separation of concerns. Moreover, being Web-based, the user interface is cross-platform, easily deployable, and accessible without the need of installing any software on the end-user machines.

5.2 Topic Creation: An Example of DIKW for DIRECT

Figure 4 presents the main page for the management of the topic creation process which allows the assessors to create the topics for the test collection.

The interface manages information resources which belong to different levels of the DIKW hierarchy and relates them in a meaningful way. Assessor and organizers can access the *data* stored and indexed in DIRECT in the form of collections of documents, and shown in relevance order after a search, and the *data* produced by assessors themselves, i.e. the informations about the topics, such as the title, description, and narrative, and the history of the changes made on those values. The latter, in particular, is shown as a branch of a tree where each node is related at the timestamp of the change made. DIRECT

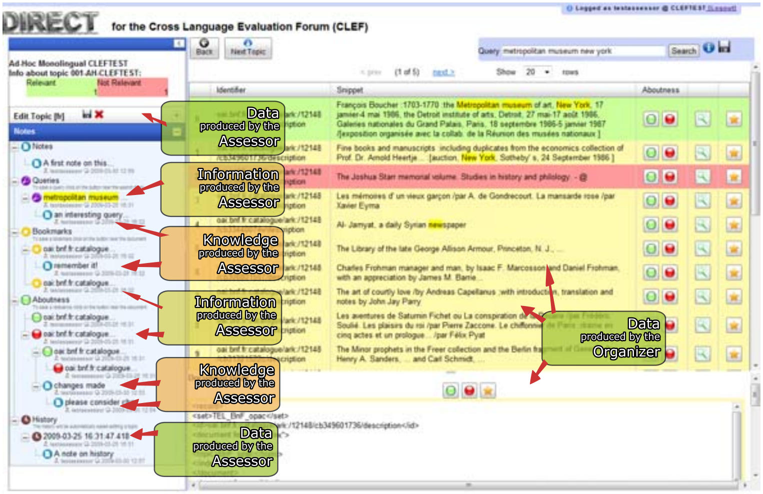
DIRECT
for the Cross Language Evaluation Forum (CLEF)
Data produced by the Assessor
Information produced by the Assessor
Knowledge produced by the Assessor
Information produced by the Assessor
Knowledge produced by the Assessor
Data produced by the Assessor
Data produced by the Organizer
Notes
Queries
Bookmarks
Aboutness
History
Back
Next Topic
Search
Identifier
Snippet
Aboutness

Fig. 4. DIRECT: creation of topics

automatically updates the tree each time a change is made, nesting the nodes related to the same topic and putting the newest near the root of the tree. This is an example of how the system can support and make explicit the creation of information resources at the *data* layer without forcing the user to taking care of the details.

You can also see how *information* and *knowledge* are produced by assessors who can save the queries used to create the topic, bookmark specific documents relevant to the topic, and save an aboutness judgement about a document in relation to the current topic. All these information resources are *information*, creating relational connections between documents and topics. Notes, comments, and discussion made by assessors are instead *knowledge*, which is created over the previous *information* and articulates into a language, and can also be attached to queries, bookmarks, and aboutness judgments.

In addition to easing the topic creation task, all these information resources are then available for conducting experiments and gaining qualitative and quantitative evidence about the pros and cons of different strategies for creating experimental collections and, thus, contribute to the advancement of the research in the field.

Finally, the possibility of interleaving and nesting different items in the hierarchy together with the ability of capturing and supporting the discussions among assessors represent, in concrete terms, a first step in the direction of making DIRECT a communication vehicle which acts as a kind of virtual research environment where the research about experimental evaluation can be carried out.

6 Conclusions

We have presented the next challenges for large-scale evaluation campaigns and their infrastructures and we have pointed out how they call for appropriate management of the knowledge process that they entail. In particular, we have discussed how digital library systems can play a key role in this scenarios and we have applied the DIKW hierarchy in the design and development of the DIRECT digital library system for scientific data.

Future work will concern the extension of the DIRECT system by adding advanced annotation functionalities in order to better support the cooperation and interaction among researchers, students, industrial partners and practicioners.

Acknowledgments

The authors would like to warmly thank Carol Peters, coordinator of CLEF and Project Coordinator of TrebleCLEF, for her continuous support and advice. The authors would like to thank Maristella Agosti and Giorgio Maria Di Nunzio for the useful discussions on the topics addressed in this chapter.

The work reported has been partially supported by the TrebleCLEF Coordination Action, as part of the Seventh Framework Programme of the European Commission, Theme ICT-1-4-1 Digital libraries and technology-enhanced learning (Contract 215231).

References

1. Ackoff, R.L.: From Data to Wisdom. Journal of Applied Systems Analysis 16, 3–9 (1989)
2. Agosti, M., Di Nunzio, G.M., Ferro, N.: A Proposal to Extend and Enrich the Scientific Data Curation of Evaluation Campaigns. In: Proc. 1st International Workshop on Evaluating Information Access (EVIA 2007), pp. 62–73. National Institute of Informatics, Tokyo (2007)
3. Agosti, M., Di Nunzio, G.M., Ferro, N., Harman, D., Peters, C.: The Future of Large-scale Evaluation Campaigns for Information Retrieval in Europe. In: Kovács, L., Fuhr, N., Meghini, C. (eds.) ECDL 2007. LNCS, vol. 4675, pp. 509–512. Springer, Heidelberg (2007)
4. Agosti, M., Ferro, N.: A Formal Model of Annotations of Digital Content. ACM Transactions on Information Systems (TOIS) 26(1), 3:1–3:57 (2008)
5. Agosti, M., Ferro, N.: Towards an Evaluation Infrastructure for DL Performance Evaluation. In: Evaluation of Digital Libraries: An Insight to Useful Applications and Methods. Chandos Publishing (2009)
6. Cleverdon, C.W.: The Cranfield Tests on Index Languages Devices. In: Readings in Information Retrieval, pp. 47–60. Morgan Kaufmann Publisher, San Francisco (1997)
7. Di Nunzio, G.M., Ferro, N.: DIRECT: a System for Evaluating Information Access Components of Digital Libraries. In: Rauber, A., Christodoulakis, S., Tjoa, A.M. (eds.) ECDL 2005. LNCS, vol. 3652, pp. 483–484. Springer, Heidelberg (2005)
8. Dussin, M., Ferro, N.: Design of a Digital Library System for Large-Scale Evaluation Campaigns. In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 400–401. Springer, Heidelberg (2008)

9. Dussin, M., Ferro, N.: The Role of the DIKW Hierarchy in the Design of a Digital Library System for the Scientific Data of Large-Scale Evaluation Campaigns. In: Proc. 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008), p. 450. ACM Press, New York (2008)
10. European Commission Information Society and Media. i2010: Digital Libraries (October 2006), http://europa.eu.int/information_society/activities/digital_libraries/doc/brochures/dl_brochure_2006.pdf
11. Ferro, N.: Digital Annotations: a Formal Model and its Applications. In: Information Access through Search Engines and Digital Libraries, pp. 113–146. Springer, Heidelberg (2008)
12. Ferro, N., Peters, C.: From CLEF to TrebleCLEF: the Evolution of the Cross-Language Evaluation Forum. In: Proc. 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, pp. 577–593. National Institute of Informatics, Tokyo (2008)
13. Fricke, M.: The Knowledge Pyramid: a Critique of the DIKW Hierarchy. Journal of Information Science 35(2), 131–142 (2009)
14. Fuhr, N., et al.: Evaluation of Digital Libraries. International Journal on Digital Libraries, 8(1):21–38 (2007)
15. Harman, D.K., Voorhess, E.M. (eds.): TREC. Experiment and Evaluation in Information Retrieval. MIT Press, Cambridge (2005)
16. National Science Board. Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century (NSB-05-40). National Science Foundation (NSF) (September 2005), http://www.nsf.gov/pubs/2005/nsb0540/
17. Nonaka, I., Takeuchi, H.: The knowledge-creating company: How Japanese companies create the dynamics of innovation. Oxford University Press, USA (1995)
18. Rowley, J.: The Wisdom Hierarchy: Representations of the DIKW Hierarchy. Journal of Information Science 33(2), 163–180 (2007)
19. Working Group on Data for Science. FROM DATA TO WISDOM: Pathways to Successful Data Management for Australian Science. Report to the Primw Minister's Science, Engineering and Innovation Council (PMSEIC) (September 2006), http://www.dest.gov.au/sectors/science_innovation/publications_resources/profiles/Presentation_Data_for_Science.htm
20. Zeleny, M.: Management Support Systems: Towards Integrated Knowledge Management. Human Systems Management 7(1), 59–70 (1987)