10 Years of CLEF Data in DIRECT: Where We Are and Where We Can Go

Maristella Agosti
agosti@dei.unipd.it

Giorgio Maria Di Nunzio
dinunzio@dei.unipd.it

Marco Dussin
dussinma@dei.unipd.it

Nicola Ferro
ferro@dei.unipd.it

Department of Information Engineering
University of Padova
Via Gradenigo, 6/b – 35131 Padova, Italy

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.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering, Relevance feedback, Retrieval models, Search process*; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*Systems issues, User issues*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Graphical user interfaces*

General Terms

Algorithms, Design, Experimentation

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, the *NII-NACSIS Test Collection for IR Systems (NTCIR)*[3] in Japan and Asia, and *INitiative for the Evaluation of XML Retrieval (INEX)*[4] now in Australia and New Zealand. Moreover, new initiatives are growing to support emerging communities and address specific issues, such as the *Forum for Information Retrieval and Evaluation (FIRE)*[5] in India.

The technologies, the services and even the users of information access systems have constantly evolved through the years with many new factors and trends influencing the field. The expectations and habits of users change 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). These new perspectives lead to the understanding 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 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 [9]. As a consequence, we have to further advance the evaluation methodologies in order to deal with the increasing complexity of the tasks. Part of this effort concerns also increasing the number of conducted experiments to have a deeper sampling and knowledge about the behaviour of such information access systems [11] as well as developing better means and tools to easily analyse, mine, and compare against a constantly increasing experimental base.

Indeed, the complexity of the tasks and the interactions to be studied and evaluated produces, as usual, valuable scientific data, which 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, communi-

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

cation, 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 good attempt in this direction is represented by the *Reliable Information Access (RIA)* Workshop [13, 14], organized by the *National Institute of Standards and Technology (NIST)* in 2003, where an in-depth study and failure analysis of the conducted experiments was performed and valuable information about them was collected [17]. However, the existence of a commonly agreed conceptual model and metadata schemas would have helped in defining and gathering the information to be kept. Similar considerations hold also for the performance measurements, the descriptive statistics, and the statistical analyses that are not explicitly modelled and for which no metadata schema is defined. It would be useful to define at least the metadata that are necessary to describe which software and which version of the software were used to compute a performance measure, which relevance judgements were used to compute a performance measure, and when the performance measure was computed. Similar metadata could be useful also for descriptive statistics and statistical analyses.

All this additional information can provide useful hints about the system models and also the context of the evaluation. The context is not simply the track or specific experiments as potentially we could need more information such as who the assessors were, how they assessed documents, what the aims of the experiment were and the circumstances in which the collection was built. Similarly, systems are more than simply a system configuration but an overall approach for a retrieval task. Furthermore, this additional information can be used to support the higher-level research activities, such as assessing the reliability of information retrieval experiments [19].

The paper introduces the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)*, the system we have developed in CLEF since 2005 to manage all the aspects of an evaluation campaign and to specifically address the issues discussed above. Section 2 gives an outlook of some of design choices and some of the functionalities currently offered by DIRECT. Then, Section 3 presents new developments currently ongoing in DIRECT and discuss some possible future directions that can stem from these novelties. Finally, Section 4 draws some conclusions.

2. THE DIRECT SYSTEM

To design and develop the DIRECT system, we approached and studied the information space entailed by an evaluation campaign in the light of the *Data, Information, Knowledge, Wisdom (DIKW)* hierarchy [1, 3, 16, 12, 18], used as a model to organize the information resources produced during it. 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* [15]), 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* [15]).
- 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.

The study contributed 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. According to the DIKW hierarchy, we can consider 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.

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

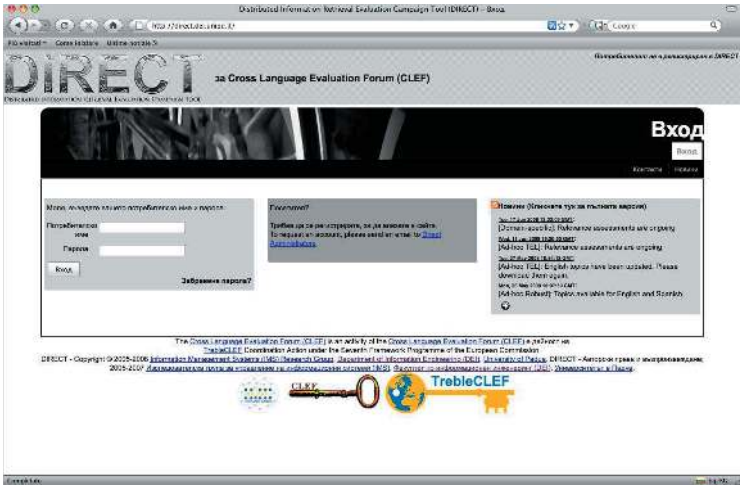

Figure 1: Bulgarian login page of DIRECT.

Figure 2: Farsi login page of DIRECT.

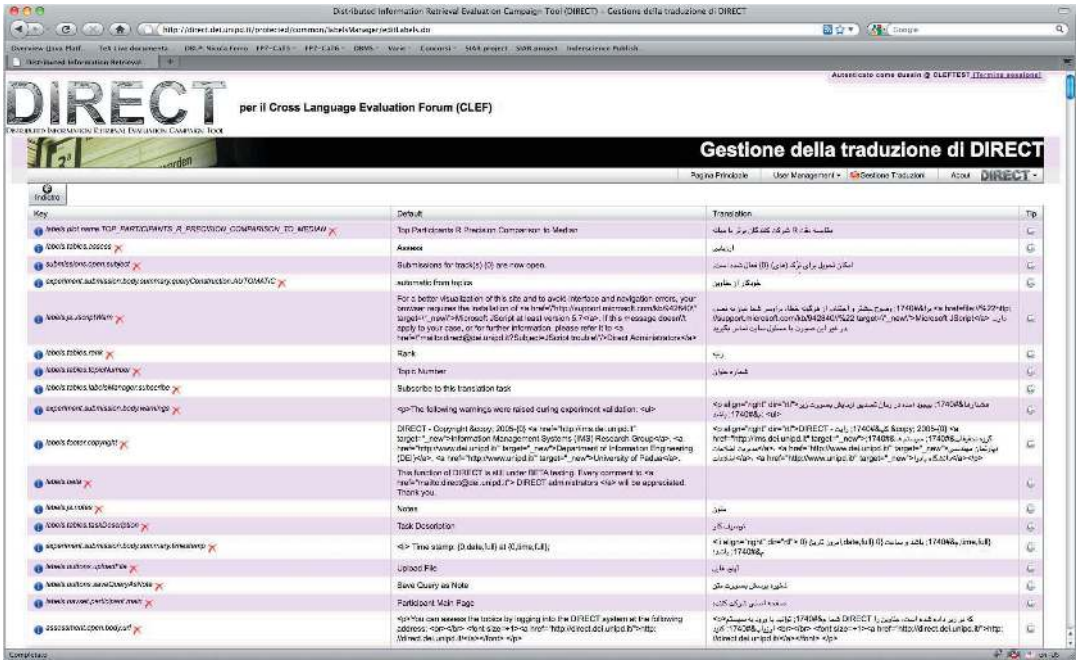

Figure 3: Internationalization manager for English to Farsi.

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, as shown in Figures 1 and 2. Moreover, Figure 3 shows the interface that users can access to localize the interface of DIRECT in their own language; currently, ten languages are supported in DIRECT: Bulgarian, Czech, English, Farsi, French, German, Indonesian, Italian, Spanish, and Portuguese.

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;
- CLEF 2009: manage 428 experiments submitted by 42 participants spread over 21 nations and assess more than 190,000 documents in four different languages, including Farsi, thanks to the work of 55 assessors.

As it will be explained further in Section 3, for CLEF 2010 DIRECT will play a different role. Indeed, CLEF 2010[6] is the continuation of the popular CLEF campaigns that have run for the past ten years. It will cover a broad range of issues in the fields of multilingual and multimodal information access evaluation. It will consist of two main parts: a peer-reviewed conference on experimental evaluation, which will innovate the CLEF tradition, and a series of laboratories, which will continue the CLEF tradition of community-based evaluation and discussion on evaluation issues. In this new setting, DIRECT will be mainly used as a support for the conference part of the whole event.

2.1 Example of Topic Creation

We applied the DIKW approach, for example, to the creation of new topics, as shown in Figure 4. 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 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 possibily 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.

Note that aboutness judgments are not definitive relevance judgments and might be imprecise since, as the topic creation task proceeds, assessor may change and evolve the topic and thus a document initially estimated as relevant might be not relevant in fact. Nevertheless these aboutness judgments represent a valuable information since they are the result of a manual and iterative search process which, in a sense, resembles the *Interactive Searching and Judging (ISJ)* stragety [6]. Therefore, we keep trace of this information and, at pooling time, we add to the pools created from the submitted runs these aboutness judgments, without informing the assessors; in this way, these potentially relevant documents get actually assessed and can effectively contribute to the robustness of the pools.

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.

3. FUTURE DIRECT(IONS)

In 2010, DIRECT will be used as the main source for dissemination of the scientific data produced during ten years of CLEF campaigns, for some core tracks, namely Ad-hoc, Domain-specific, and GeoCLEF. We are also experimenting the support for multimedia tracks by making available data of some ImageCLEF tasks.

In this way, both participants of previous CLEF editions and researchers who have never participated will be able to access all the data available: topics, pools, experiments, statistical measures, plots, and so on. Moreover, this will provide support for the CLEF 2010 conference since it will provide researchers and interested parties with a coherent

[6] http://www.clef2010.org/

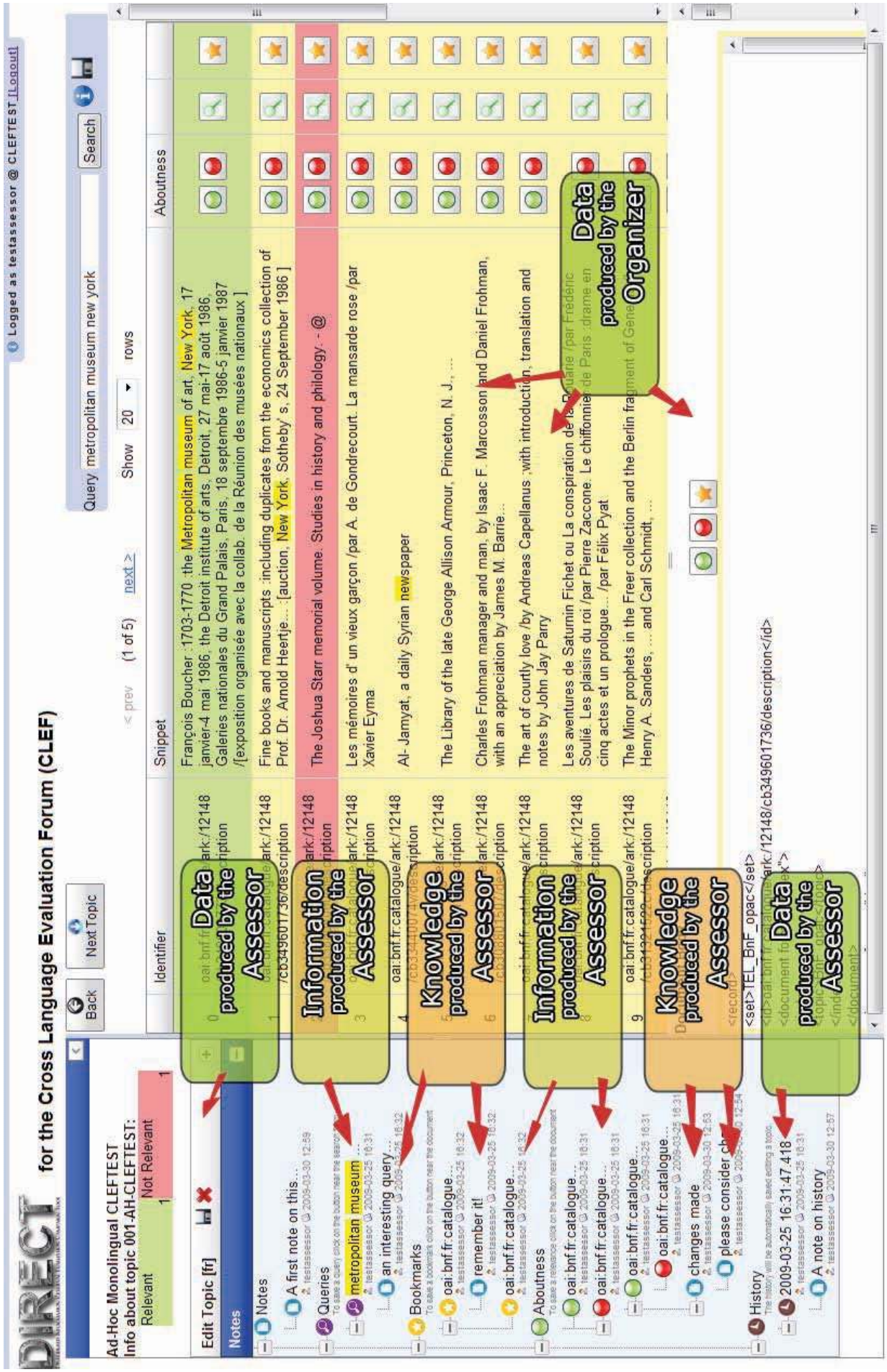

Figure 4: DIRECT: creation of topics.

Table 1: Summary of the data available in DIRECT.

Campaign	Tracks	Tasks	Pools (assessments)	Topics (languages)	Experiments (participants)
CLEF 2000	2	5	5 (91,408)	366 (8)	95 (20)
CLEF 2001	2	10	8 (181,884)	675 (12)	192 (31)
CLEF 2002	2	20	10 (242,019)	672 (12)	282 (33)
CLEF 2003	2	21	13 (445,757)	711 (11)	415 (33)
CLEF 2004	2	15	8 (207,171)	725 (13)	283 (26)
CLEF 2005	3	22	12 (372,893)	1,097 (16)	465 (33)
CLEF 2006	4	35	20 (665,801)	2,141 (19)	455 (37)
CLEF 2007	4	25	14 (303,057)	2,698 (23)	409 (34)
CLEF 2008	5	25	12 (302,295)	2,261 (20)	482 (36)
CLEF 2009	5	18	10 (351,023)	3,119 (21)	422 (41)
TOTALS	–	–	112 (3,163,308)	14,465 (–)	3,500 (–)

and online access to ten years of data and give them the possibility of conducting longitudinal studies over the CLEF data as well as deeply mine and analyze them. Up to now, about 40 new users around the world have registerd to get access to these data.

Table 1 summarizes all the data that can be accessed through the DIRECT system and which are now available to the research and developer communities.

Figure 5 shows a screenshot of the interface for accessing the whole set of scientific data produced during the history of CLEF. On the left, there is a tree which allows the user to browse thorough the CLEF campaigns from 2000 to 2009 and, for each campaign, it is possible to see what tracks and tasks are available.

Once the user has selected a task, he is presented, in the right pane, with the list of experiments submitted for that task plus specific information about each experiment, such as a description, the source language, which fields of the topic have been used to construct the query, whether it is an automatic or manual experiment, whether it has been pooled or not, and so on. At this point, the user can decide to view and access metrics and performance measures about the experiment, as shown in Figure 6 or to directly download it for further re-use and analysis. It is also possible to select multiple experiments and download them at once. In the left pane, once a task has been selected, it is possible either to download the relevance assessments and pool corresponding to that task, as well as the topics used in the task, in multiple languages if available and visualize overall statistics about the task, , as shown in Figure 7.

This new functionality of the system represents a first step in the direction of being able to easily and systematically carry out longitudinal studies to assess the improvement in the performances over the years [2] and to better understand systems behaviour with respect to languages and tasks. As it emerges from the discussion in Section 1, this will become a more and more compelling need to be able to evaluate and study increasingly complex information access systems.

Moreover, the gathered data and the possibility of seamlessly access them will foster the re-use of experimental collections and data, for example, to understand how new technologies apply and perform over time. This has been pioneered in CLEF 2005 when the "Ad-hoc Two Years On" task has been offered [8]. The objective of the tasks was to re-assess truly multilingual information access: participants have been asked to re-use existing experiments and collections to understand what was the difference between better merging existing multilingual runs and applying new systems to the same document collections.

In the perspective of the upcoming challenges previously discussed, our final goal is to turn DIRECT [5] from a *Digital Library System (DLS)* 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 and this can be exploited to improve the exploitation and the comparison to the existing experimental base. Thus, we aim at making DIRECT an active communication vehicle for the communities interested in the experimental evaluation by extending it 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 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, 10] 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.

The future work will concern to turn this kind of "repository" and knowledge-base of experimental data into an active research tool by applying the annotation, bookmarking, querying, and discussion facilities we have already developed for the creation of topics. In this way, users – students, researchers, developers, and practicioners – will be able to exploit the interface for accessing the experimental data to enrich the managed experiments, measures, pools, and topics with their own annotations and content and share their ideas and thoughts about such data. In this scenario, it will become, for example, possible to online discuss and compare experiments, add explanations about why an experiment performs better or worse than another one, link this

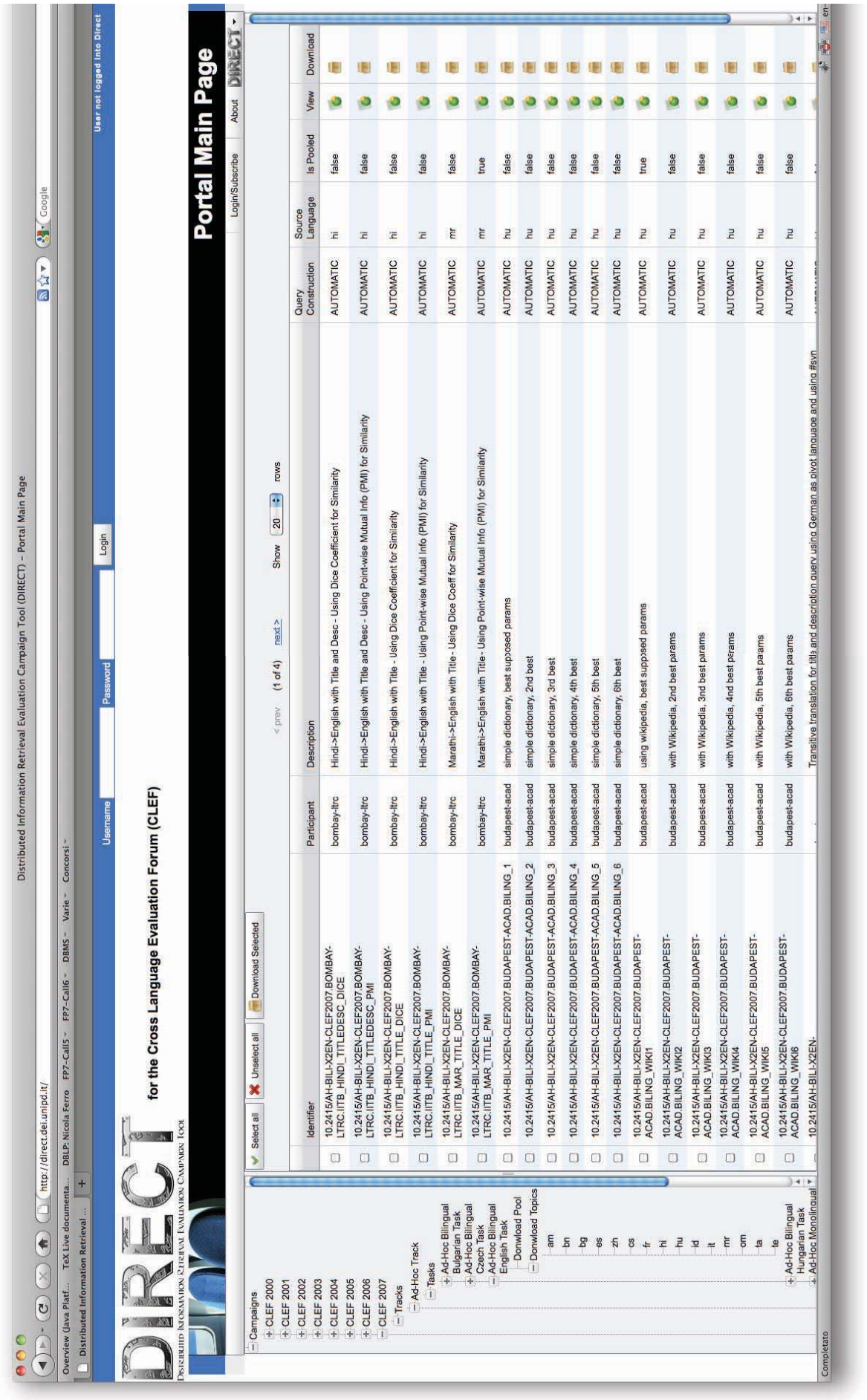

Figure 5: DIRECT interface for accessing the history of ten years of CLEF data.

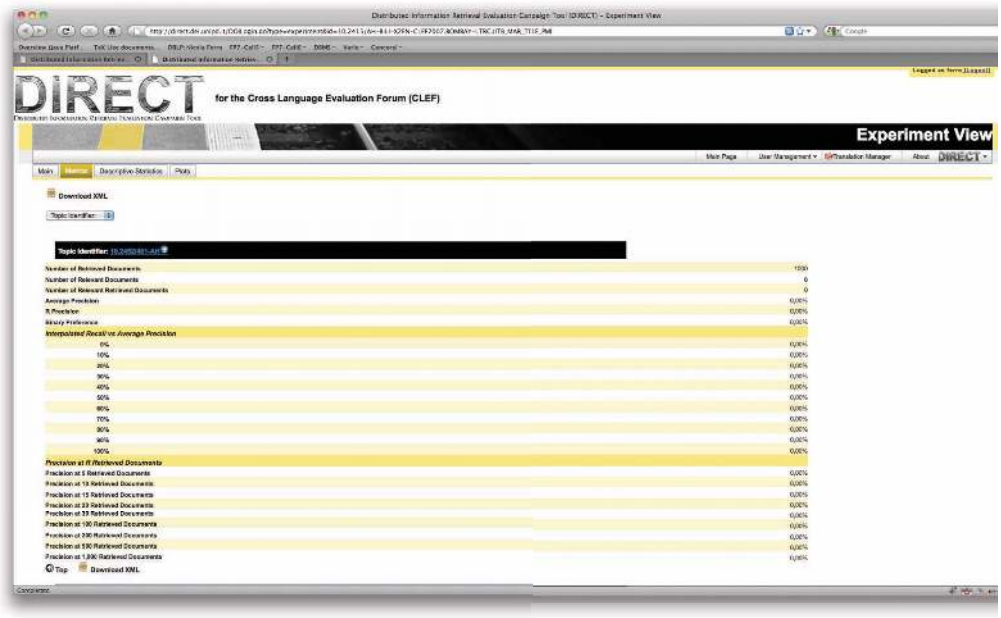

Figure 6: DIRECT interface for accessing performance measures about an experiment.

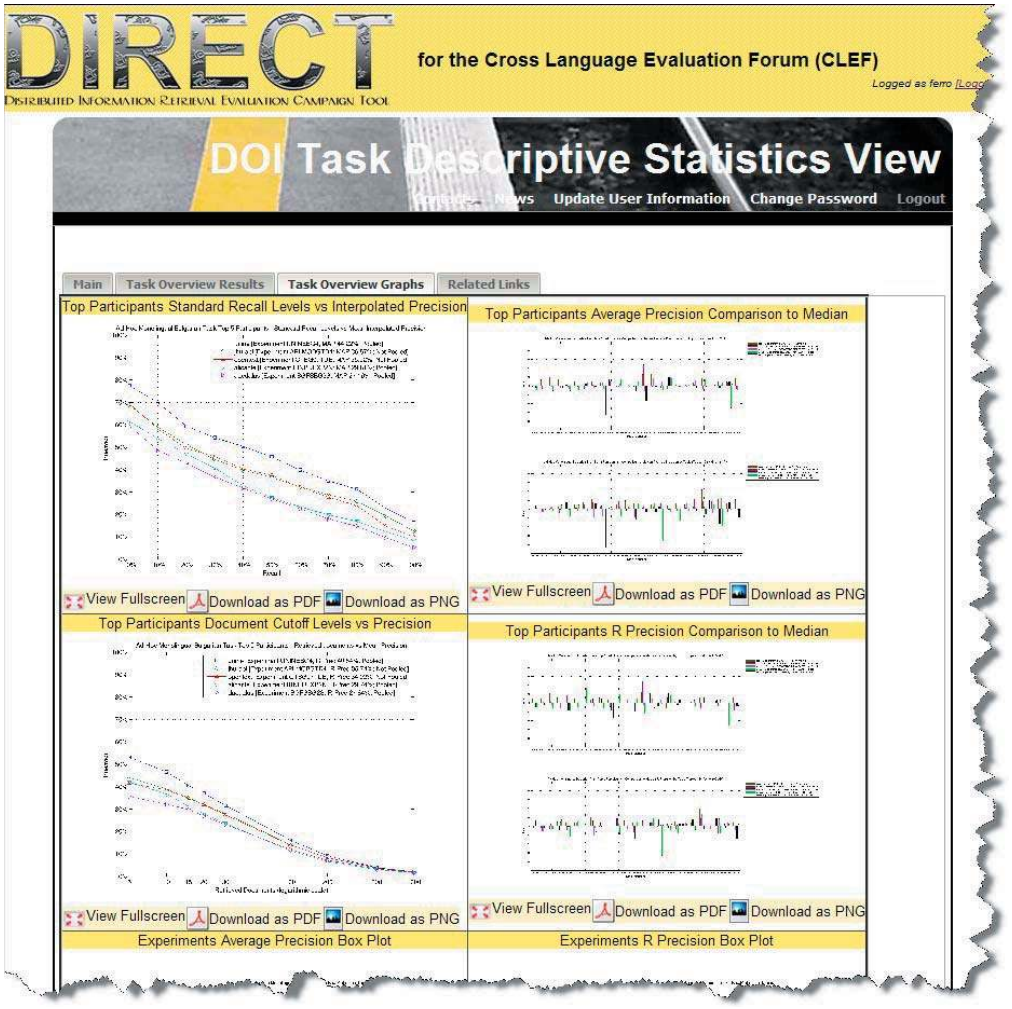

Figure 7: DIRECT interface for accessing plots and descriptive statistics about a task.

experiment to external (Web) resource pertinent to it, and so on.

4. CONCLUSIONS

The paper has discussed some compelling issues that large-scale evaluation campaigns should take into consideration when they come to the management, description, and access to the scientific data produced during their course.

We have then presented the DIRECT system, we have developed in CLEF since 2005 in order to start to address some of those issues.

Finally, we have discussed some ongoing activities targeted towards using DIRECT not only as a campaign management tool but also as a dissemination source for the scientific data produced during the last ten years of CLEF campaigns. Moreover, we have outlined some possible future directions that we will pursue to favour an active involvement of the users with the managed data.

Much work is still to come, such as for example to conduct user studies to assess the actual utilization of the system and to gather suggestions for possible future directions.

5. ACKNOWLEDGMENTS

The authors would like to warmly thank Carol Peters for her continuous support and advice.

6. REFERENCES

[1] R. L. Ackoff. From Data to Wisdom. *Journal of Applied Systems Analysis*, 16:3–9, 1989.

[2] M. Agosti, G. M. Di Nunzio, and N. Ferro. Evaluation of a Digital Library System. In M. Agosti and N. Fuhr, editors, *Notes of the DELOS WP7 Workshop on the Evaluation of Digital Libraries*, pages 73–78. `http://dlib.ionio.gr/wp7/workshop2004_program.html`, 2004.

[3] M. Agosti, G. M. Di Nunzio, and N. Ferro. A Proposal to Extend and Enrich the Scientific Data Curation of Evaluation Campaigns. In T. Sakay, M. Sanderson, and D. K. Evans, editors, *Proc. 1st International Workshop on Evaluating Information Access (EVIA 2007)*, pages 62–73. National Institute of Informatics, Tokyo, Japan, 2007.

[4] M. Agosti and N. Ferro. A Formal Model of Annotations of Digital Content. *ACM Transactions on Information Systems (TOIS)*, 26(1):3:1–3:57, 2008.

[5] M. Agosti and N. Ferro. Towards an Evaluation Infrastructure for DL Performance Evaluation. In G. Tsakonas and C. Papatheodorou, editors, *Evaluation of Digital Libraries: An insight into useful applications and methods*, pages 93–120. Chandos Publishing, Oxford, UK, 2009.

[6] G. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient Construction of Large Test Collections. In Croft et al. [7], pages 282–289.

[7] W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors. *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*. ACM Press, New York, USA, 1998.

[8] G. M. Di Nunzio, N. Ferro, G. J. F. Jones, and C. Peters. CLEF 2005: Ad Hoc Track Overview. In C. Peters, F. C. Gey, J. Gonzalo, G. J. F. Jones, M. Kluck, B. Magnini, H. Müller, and M. de Rijke, editors, *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross–Language Evaluation Forum (CLEF 2005). Revised Selected Papers*, pages 11–36. Lecture Notes in Computer Science (LNCS) 4022, Springer, Heidelberg, Germany, 2006.

[9] M. Dussin and N. Ferro. Managing the Knowledge Creation Process of Large-Scale Evaluation Campaigns. In M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, and G. Tsakonas, editors, *Proc. 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*, pages 63–74. Lecture Notes in Computer Science (LNCS) 5714, Springer, Heidelberg, Germany, 2009.

[10] N. Ferro. Digital Annotations: a Formal Model and its Applications. In M. Agosti, editor, *Information Access through Search Engines and Digital Libraries*, pages 113–146. Springer-Verlag, Heidelberg, Germany, 2008.

[11] N. Ferro and D. Harman. CLEF 2009: Grid@CLEF Pilot Track Overview. In C. Peters, G. M. Di Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Peñas, and G. Roda, editors, *Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments – Tenth Workshop of the Cross–Language Evaluation Forum (CLEF 2009). Revised Selected Papers*. Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany, 2010.

[12] M. Fricke. The Knowledge Pyramid: a Critique of the DIKW Hierarchy. *Journal of Information Science*, 35(2):131–142, 2009.

[13] D. Harman and C. Buckley. The NRRC Reliable Information Access (RIA) Workshop. In M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, editors, *Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 528–529. ACM Press, New York, USA, 2004.

[14] D. Harman and C. Buckley. Overview of the Reliable Information Access Workshop. *Information Retrieval*, 12(6):615–641, 2009.

[15] I. Nonaka and H. Takeuchi. *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. Oxford University Press, USA, 1995.

[16] J. Rowley. The Wisdom Hierarchy: Representations of the DIKW Hierarchy. *Journal of Information Science*, 33(2):163–180, 2007.

[17] I. Soboroff. A guide to the RIA workshop data archive. *Information Retrieval*, 12(6):642–651, 2009.

[18] M. Zeleny. Management Support Systems: Towards Integrated Knowledge Management. *Human Systems Management*, 7(1):59–70, 1987.

[19] J. Zobel. How Reliable are the Results of Large-Scale Information Retrieval Experiments. In Croft et al. [7], pages 307–314.