Project no. 215231

TrebleCLEF

Treble-CLEF: Evaluation, Best Practices and Collaboration for Multilingual Information Access
IST: ICT-1-4-1, Digital libraries and technology-enhanced learning

Deliverable 2.2
Operational Scientific Digital Library

Start Date of Project: 01 January 2008
Duration: 24 Months

Organisation Name of Lead Contractor for this Deliverable: UNIPD

Version 1.1, July 2008 [final]

Project co-funded by the European Commission within the Seventh Framework programme

Document Information

Deliverable number:	2.2
Deliverable title:	Operational Scientific Digital Library
Due date of deliverable:	30/06/2008
Actual date of deliverable:	18/07/2008
Author(s):	Nicola Ferro, UNIPD
Participant(s):	UNIPD
Workpackage:	2
Workpackage title:	Evaluation Infrastructure
Workpackage leader:	UNIPD
Dissemination Level:	PU
Version:	1.1
Keywords:	large-scale evaluation campaigns, scientific data, digital library system, DIRECT

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
1.0	26/06/2008	Draft	UNIPD	Circulated to all partners
1.1	18/07/2008	Final	UNIPD	Revised after partners' comments

Abstract

The data produced during an evaluation campaign of information access systems are valuable scientific data, and as a consequence, it is important to keep trace of their lineage since this allows us to judge their quality and applicability for a given use. As the data is frequently enriched by the addition of further analyses and interpretations, it is necessary to provide an effective way be able to cite both the data and any additional processing. This document describes the design and development of a digital library system for scientific data, called DIRECT, which not only manages the different types of information resources employed in a large-scale evaluation campaign and supports the different stages of the campaign, but also facilitates the sharing and dissemination of the results. DIRECT is proving an important instrument to improve cooperation among researchers and to facilitate the transfer of scientific and innovative results.

Table of Contents

Executive Summary

Scientific data, their management, access and reuse through citation, curation, enrichment, and preservation are essential components of scientific research. We consider information retrieval experimental evaluation as a source of valuable scientific data and, in this context, we propose a data curation approach as an extension to the traditional methodology in order to better manage, preserve, interpret, and enrich the scientific data produced and to effectively promote the transfer of knowledge.

The current approach to experimental evaluation is mainly focused on creating comparable experiments and evaluating their performance whereas researchers would also greatly benefit from an integrated vision of the scientific data produced, together with analyses and interpretations, and from the possibility of keeping, re-using, and enriching them with further information. The way in which experimental results are managed, made accessible, exchanged, visualized, interpreted, enriched and referenced is an integral part of the process of knowledge transfer and sharing towards relevant application communities, such as the digital library one.

Therefore, we have designed and developed a digital library system for scientific data able to support the course of an evaluation initiative and to promote the dissemination and sharing of the experimental results. In order to achieve this goal, we had to:

- analyse the different needs of the actors involved in the evaluation campaign and provide an access strategy to the relevant information resources tailored to their needs;
- introduce a conceptual model clearly defining the entities implied by the information space, their features and their relationships;
- develop common metadata formats, which provide meaning to the data, enable their sharing and reuse, and keep track of their lineage;
- adopt a unique identification mechanism, which allows the citation of and easy access to the scientific data and supports their enrichment;
- manage all the aspects of the campaign, such as track set-up, management of document collections, topic creation, experiment submission, relevance assessments, computation of statistical analyses, visualization of and access to the scientific data managed, data exchange, and so on.

The result of our work is the Distributed Information Retrieval Evaluation Campaign Tool (DIRECT) [1], a digital library system for managing the scientific data and information resources produced during an evaluation campaign. A preliminary version of DIRECT was introduced into CLEF in 2005 and subsequently tested and developed in the CLEF 2006 and 2007 campaigns. In this deliverable we present a thoroughly revised and considerably extended version which has been developed and implemented within the context of TrebleCLEF.

In the ongoing CLEF 2008 campaign, DIRECT is being used by over 130 participants spread over 20 different nations, who have submitted 490 experiments. Within the DIRECT framework, 80 assessors from more than ten countries have created over 200 topics in seven different languages and are assessing about 250,000 documents, including document in languages like Russian, which uses the Cyrillic alphabet, and Farsi, which is written from right to left.

DIRECT can be exploited as an effective tool to foster knowledge transfer towards not only the IR research community but also relevant application communities and the industry. Indeed, currently, it manages the information resources produced during the CLEF campaigns which are especially interesting for the IR research community. Nevertheless, when used in the context of other TrebleCLEF tasks, such as “Task 3.1 Best practices in system-oriented aspects of MLIA applications” and “Task 4.4 Grid Experiments”, DIRECT can manage the experimental results needed to support and explain the indications and guidelines which are the outcomes of these tasks. Indeed, for example, the guidelines could provide hints about what is the best stemmer to be used for a given language and give a reference to the experiment and performance measures, managed by DIRECT, which support the assertion. In this way, system and application developers not only benefit from the findings

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

summarized in the guidelines and best practices but also can have an idea of the actual performances they can expect adopting a solution or another by accessing the only the information relevant to them through DIRECT.

This document describes the design and development of the DIRECT digital library system. In particular, it discusses the conceptual framework proposed for modeling the information space of an evaluation campaign. It then presents an analysis of the user requirements for the different stakeholders involved in an evaluation campaign and describes the main contributions and achievements of the digital library system developed. In addition, it provides details on the architecture of the DIRECT system and illustrates some of the functionality offered by the current prototype.

1 Introduction

International large-scale evaluation campaigns for IR, such as the Text REtrieval Conference (TREC)[2], the Cross-Language Evaluation Forum (CLEF)[3] and the NII-NACSIS Test Collection for IR Systems (NTCIR)[4], promote and stimulate the research and development of Information Retrieval Systems (IRS) by:

- creating an evaluation infrastructure and organizing regular evaluation campaigns for system testing where ideas can be exchanged and different approaches can be discussed;
- building a strong multidisciplinary research community where problems are faced from different points of view - e.g. information retrieval, question answering, natural language processing - and multiple techniques are merged and harmonized together;
- constructing publicly available test-suites that can also be used outside the evaluation campaigns for system benchmarking.

Furthermore, large-scale evaluation campaigns impact not only on the Information Retrieval (IR) field but also on other research fields, which adopt and apply the results, such as the Digital Library (DL) field. The information access and extraction components of a Digital Library System (DLS), which index, search and retrieve documents in response to a user's query, rely on methods and techniques taken from IR. 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 DLS [22].

During the workshop on "The Future of Large-scale Evaluation Campaigns"[5] [6], which was organised jointly by the University of Padua and the DELOS Network of Excellence on Digital Libraries[6] and held in Padua, Italy, March 2007, a critical assessment of the scientific results of the CLEF initiative was conducted and recommendations were made about the key research areas to pursue in the near future:

- *user modelling*, e.g. what are the requirements of different classes of users when querying multilingual information sources;
- *language-specific experimentation*, e.g. looking at differences across languages in order to derive best practices for each language, best practices for component development and best practices for MLIA systems as a whole;
- *results presentation*, e.g. how can results be presented in the most useful and comprehensible way to the user;
- *performance measurements,* e.g. identifying new metrics specifically designed and tuned for use in a multilingual context, studying new methods for creating test collections quickly and efficiently;
- *experimental result management*, e.g. understanding that the experimental data produced during an evaluation campaign are valuable scientific data, and as a consequence, should be archived, enriched, and curated in order to ensure future accessibility and re-use.

In this context, TrebleCLEF intends to promote research, development, implementation and industrial take-up of multilingual, multimodal information access functionality in the following ways [10]:

- by continuing to support the annual CLEF system evaluation campaigns with tracks and tasks designed to stimulate R&D to meet the requirements of the user and application communities;
- by constituting a scientific forum for the MultiLingual Information Access (MLIA) community of researchers enabling them to meet and discuss results, emerging trends, new directions;

[2] http://trec.nist.gov/
[3] http://www.clef-campaign.org/
[4] http://research.nii.ac.jp/ntcir/index-en.html
[5] http://ims.dei.unipd.it/events/2007/future-evaluation-campaigns/future-eval-index.html
[6] http://www.delos.info/

- by acting as a virtual centre of competence providing a central reference point for anyone interested in studying or implementing MLIA functionality and encouraging the dissemination of information.

Moreover, TrebleCLEF aims at fostering and supporting the transfer of knowledge about MLIA towards the DL community especially in the context of the i2010 Digital Library Initiative, which clearly states that the improvement of multilingual and multicultural information access and search is one of the key objectives necessary to provide access to quality digital content for all. This is also a key factor in the process towards the European Digital Library[7] where multilinguality and MLIA are seen as issues of interoperability over the different languages managed by the national systems, which will be part of the European Digital Library [23].

The growing interest in the correct management of scientific data has been recently highlighted by different organizations, among them the European Commission, the US National Scientific Board, and the Australian Working Group on Data for Science.

The European Commission in the i2010 Digital Library Initiative clearly states that "digital repositories of scientific information are essential elements to build European eInfrastructure for knowledge sharing and transfer, feeding the cycles of scientific research and innovation up-take" [20]. The US National Scientific Board points out that "organizations make choices on behalf of the current and future user community on issues such as collection access; collection structure; technical standards and processes for data curation; ontology development; annotation; and peer review". And, those organizations "are uniquely positioned to take leadership roles in developing a comprehensive strategy for long-lived digital data collections" [31]. The Australian Working Group on Data for Science suggests to "establish a nationally supported long-term strategic framework for scientific data management, including guiding principles, policies, best practices and infrastructure", that "standards and standards-based technologies be adopted and that their use be widely promoted to ensure interoperability between data, metadata, and data management systems", and that "the principle of open equitable access to publicly-funded scientific data be adopted wherever possible [. . .] As part of this strategy, and to enable current and future data and information resources to be shared, mechanisms to enable the discovery of, and access to, data and information resources must be encouraged" [36].

These observations suggest that considering the IR experimental evaluation as a source of scientific data implies not only rethinking the evaluation methodology employed but also re-considering the way in which evaluation campaigns are organized. Changes to the IR evaluation methodology need to be correctly supported by organizational, hardware, and software infrastructures which allow for management, search, access, curation, enrichment, and citation of the produced scientific data.

Such change also involves the organizations, which run the evaluation campaigns, since they have not only to provide the infrastructure but also to participate in the design and development of it. As highlighted by [31], these organizations should take a leading role in developing a comprehensive strategy for long-lived digital data collections and drive the research community through this process in order to improve the way of doing research. As a consequence, the aim and the reach of an evaluation campaign would be widened because, besides bringing research groups together and providing them with the means for discussing and comparing their work, the campaign will also define guiding principles, policies, best practices for making use of the scientific data produced.

We have taken all the above-mentioned points into account when designing an evaluation campaign, which has two main objectives: first, it should promote research in the IR field by highlighting valuable areas which need to be explored and by offering the means for conducting, comparing, and discussing experiments. Second, it should make the management and curation of the scientific data produced an integral part of the IR research process. Therefore, in our vision, an evaluation campaign has to provide guidelines, best practices, conceptual and logical models for data representation and exchange, preservation and curation of the scientific data.

[7] http://www.europeana.eu/

As a consequence, an evaluation campaign has to provide a software infrastructure suitable for carrying out this second new role. A digital library system seems to be the natural choice for managing and enriching all the information resources produced during an evaluation campaign. Information enrichment is one of the main activities supported by a DLS. Of particular relevance, we can cite provenance as "important in judging the quality and applicability of information for a given use and for determining when changes at sources require revising derived information" [28] and citation, intended as the possibility of explicitly mentioning and making references to portions of a given digital object.

The result of our work is Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)[8], a DLS for managing the scientific data and information resources produced during an evaluation campaign. A preliminary version of DIRECT was introduced into CLEF in 2005 [14] and subsequently tested and developed in the CLEF 2006 [15] and CLEF 2007 [3] campaigns. In this deliverable we present a thoroughly revised and considerably extended version which has been developed and implemented within the context of TrebleCLEF.

The document is organized as follows: Section 2 discusses the conceptual framework proposed for modeling the information space of an evaluation campaign; Section 3 presents analyses of the user requirements of the different stakeholders involved in an evaluation campaign; Section 4 describes the main contributions and achievements of the DIRECT digital library; Section 5 provides details on the architecture of the DIRECT system; Section 6 illustrates some of the functionality offered by the current prototype; finally, Section 7 draws some conclusions.

2 Conceptual Framework for the Information Space of an Evaluation Campaign

The current approach for laboratory evaluation of information access systems relies on the Cranfield methodology, which makes use of *experimental collections* [12]. An experimental collection is a triple $C = (D, T, J)$, where: D is a set of documents, called also collection of documents; T is a set of topics, which expresses the user's information needs and from which the actual queries are derived; J is a set of relevance judgements, i.e. for each topic $t \in T$ and for each document $d \in D$ it is determined whether d is relevant to t or not.

An experimental collection C allows the comparison of information access systems according to some measurements, which quantify their performances. The main goal of an experimental collection is both to provide a common test-bed to be indexed and searched by information access systems and to guarantee the possibility of replicating the experiments.

When reasoning about this evaluation paradigm, a first step is to point out that the experimental evaluation in the IR field is a scientific activity and, as such, its outcomes are different kinds of valuable scientific data. So, the experiments themselves represent our primary scientific data and the starting point of our investigation. Using the experimental data, we produce different performance measurements, such as precision and recall which are standard measures that are used to evaluate the performances of an IRS for a given experiment. Starting from these performance measurements, we can compute descriptive statistics, such as mean or median, used to summarize the overall performances achieved by an experiment or by a collection of experiments. Finally, we can perform hypothesis tests and other statistical analyses to conduct an in-depth analysis and comparison over a set of experiments.

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

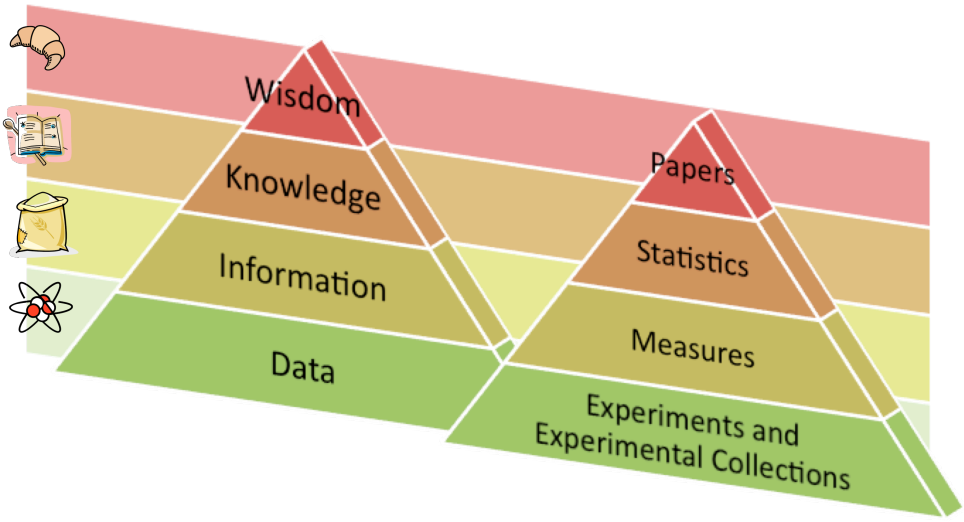

Figure 1: The DIKW hierarchy with respect to the experimental evaluation.

We can frame the above-mentioned scientific data in the context of the Data, Information, Knowledge, Wisdom (DIKW) hierarchy [2] and [38], represented in Figure 1:

- at the *data layer* there are raw, basic elements, partial and atomized, which have little meaning by themselves and no significance beyond their existence. Data are created with facts, can be measured, and can be viewed as the building blocks of the other layers. Despite the possibility of manipulation, a limited amount of actions can be performed with them;
- the *information layer* is the result of computations and processing of the data. Information comes from the form taken by the data when they are grouped and organised in different ways to create relational connections: indeed, the term "inform" itself means etymologically to give shape, to form, thus entailing the notion of giving data a new shape by relating them together and with other entities;
- 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. It can be articulated into a language, more or less formal, such as words, numbers, expressions and so on, transmitted to others, or be embedded in individual experience, like beliefs or intuitions;
- the *wisdom level* provides interpretation, explanation, and formalization of the content of the previous levels. Wisdom is not one thing: it is the highest level of understanding, and is a uniquely human state. The previous levels are related to the past, with wisdom people can strive to the future.

As an example, we can consider the task of baking bread [38]: "*data* are like basic elements: atoms and molecules of starch, H_2O, bacteria of yeast, etc.; no trace of bread anywhere. *Information* is like ingredients: flour, sugar, water, spices; still no trace of the intended outcome (but one cannot make a beer out of it anymore). Having all such ingredients does not imply that a *knowledge* of how to make bread exists: one can still end up with a tasty crust, black cinder or gluey mush. Knowledge involves relations: recipes and their contextual interpretations. Further, having the know-how for making bread does not imply that one actually should make bread and why. *Wisdom*, goes beyond knowledge because it allows comparisons (judgments) with regard to know-what and know-why. It is a long way from data to wisdom".

As shown in Figure 1, the scientific data produced during an evaluation campaign can be framed into the DIKW hierarchy:

- *data*: the *experimental collections* and the *experiments* correspond to the "data level" in the hierarchy, since they are the raw, basic elements needed for any further investigation and they would have little meaning by themselves. In fact, an experiment and the list of results obtained conducting it are almost useless without a relationship with the experimental collection with

respect to which the experiment has been conducted and the list of results produced; those data constitute the basis for any subsequent computation;

- *information*: the *performance measurements* correspond to the "information level" in the hierarchy, since they are the result of computations and processing on the data, so that we have associated a meaning with the data by way of some kind of relational connection. For example, precision and recall measures are obtained by relating the list of results contained in an experiment with the relevance judgements *J*;
- *knowledge*: the *descriptive statistics* and the *hypothesis tests* correspond to the "knowledge level" in the hierarchy, since they are a further elaboration of the information carried by the performance measurements and provide us with some insights about the experiments;
- *wisdom*: theories, models, algorithms, techniques, and observations, which are usually communicated by means of papers, talks, and seminars, correspond to the "wisdom level" in the hierarchy, since they provide interpretation, explanation, and formalization of the content of the previous levels.

As observed by [38], "while data and information (being components) can be generated per se, i.e., without direct human interpretation, knowledge and wisdom (being relations) cannot: they are human- and context-dependent and cannot be contemplated without involving human (not machine) comparison, decision making and judgement". This observation fits also the case of IR experimental evaluation. Indeed, experiments (data) and performance measurements (information) are usually generated in an automatic way by IRS, programs and tools for assessing performances. On the other hand, statistical analyses (knowledge) and models and algorithms (wisdom) require a deep involvement of researchers in order to be conducted and developed.

This view of the IR experimental evaluation calls into question whether the Cranfield methodology is able to support an experimental approach where the whole process from data to wisdom is taken into account [3].

This question is made more compelling by the fact that, when we deal with scientific data, "the lineage (provenance) of the data must be tracked, since a scientist needs to know where the data came from [. . .] and what cleaning, rescaling, or modelling was done to arrive at the data to be interpreted" [1]. Moreover, [28] points out how provenance is "important in judging the quality and applicability of information for a given use and for determining when changes at sources require revising derived information". Furthermore, when scientific data are maintained for further and future use, they should be enriched and, sometimes, the enrichment of a portion of scientific data can make use of a citation for explicitly mentioning and making references to useful information [4]. Finally, [31] highlights that "digital data collections enable analysis at unprecedented levels of accuracy and sophistication and provide novel insights through innovative information integration".

Therefore, the question turns out to be not only to which degree the Cranfield methodology embraces the passing from data to wisdom but also whether the proper strategies are adopted to ensure the provenance, the enrichment, the citation, and the interpretation of the scientific data.

3 User Requirements Analysis

Different types of actors are involved in an evaluation campaign:

- the *participant* takes part in the evaluation campaign in order to have a forum to test his new algorithms and techniques, to compare their effectiveness, and to discuss and share his proposals. He needs support for the submission of his experiments and their validation; he then expects to receive measurements about the performances of his experiments and overall indicators that allows for the comparison of his experiments with the ones submitted by other participants. Moreover, he should have the possibility of properly citing his experiments and other information resources and to get a citation correctly resolved to the corresponding information resources;

- the *assessor* contributes to the creation of the experimental collections by both proposing the topics and assessing the relevance of the documents with respect to those topics. He needs support in both these tasks which are labour-intensive and require the inspection of great amounts of data;
- the *visitor* needs to consult, browse, and access all the information resources produced during the course of an evaluation campaign in a meaningful fashion which provides him insights about the conducted experiments. Moreover, he should have the possibility of properly citing the accessed information resources and to get a citation correctly resolved to the corresponding information resources;
- the *organizer* manages the different aspects of an evaluation forum: he contributes to the creation of the experimental collections by preparing the documents and overseeing the creation of the topics and the relevance assessments; he provides the framework for the participants to conduct their experiments and for the assessors to create the topics and perform the relevance assessments; he computes the different measures for assessing the performances of the submitted experiments as well as descriptive statistics and statistical tests to characterize the overall features of the submitted experiments; finally, he provides the visitors with the means for accessing all the information resources they are looking for.

These actors interact together in various ways during the course of an evaluation campaign and contribute differently to the DIKW hierarchy discussed above.

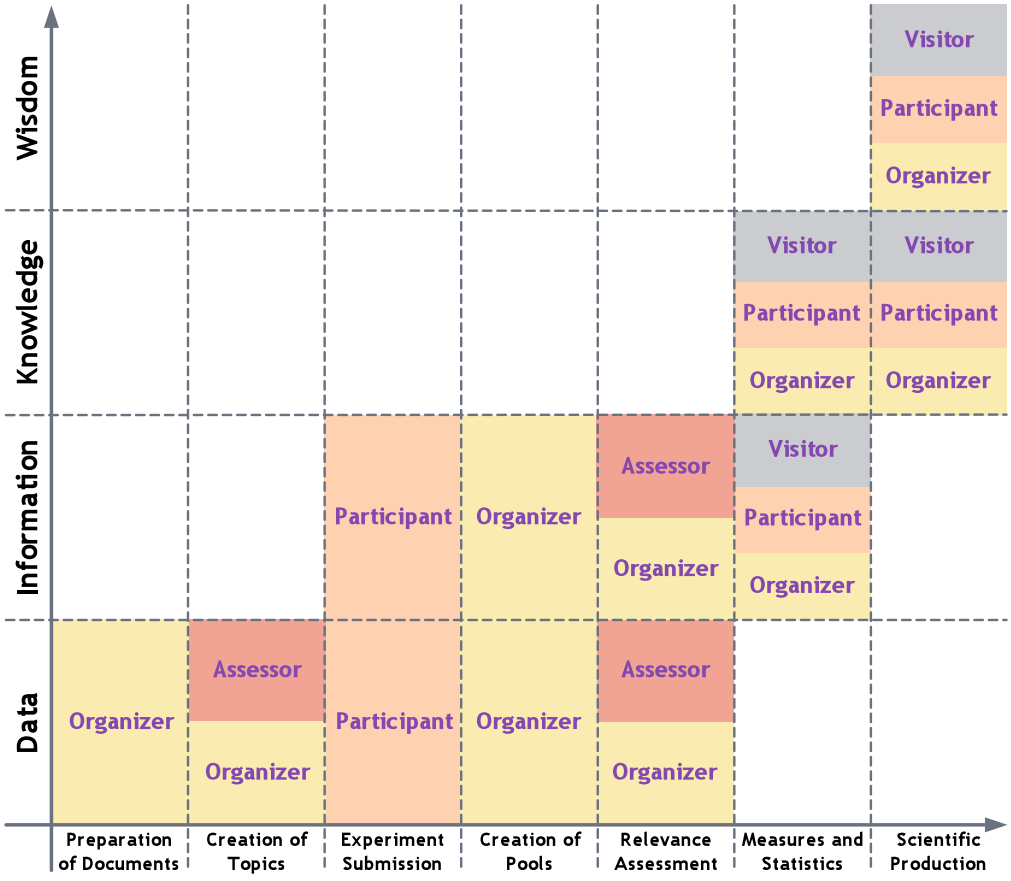

Figure 2: Relationship between the DIKW hierarchy, the different types of actors, and the main steps of an evaluation campaign.

Figure 2 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. In addition, Figure 2 shows the degree to which each step has been implemented in the user interface of the current prototype.

From Figure 2, it can be noted that the early stages of an evaluation campaign are mainly devoted to the preparation of the data and require limited interaction between the different actors. As time passes and we enter in the heart of the campaign, there is a progressive movement from data to wisdom and also the number of actors involved and their interaction grows.

We will now examine in more detail the elements shown in Figure 2:

- *acquisition and preparation of documents*: the organizers are responsible for acquiring, formatting and preparing the set of documents that will be released to the participants. These documents are part of the data on which the experiments are built on.

  Organizers need an interface that allows them to upload the collections of documents, that can be in diverse media, into the DLS in order to make them available to participants and assessors;
- *creation of topics*: the organizers and the assessors cooperate together to create the topics for the test collection. For each topic, this step usually requires preparing a first draft of the topic[9] and searching the set of document to verify that there are relevant documents for that topic; then the topic is refined by discussing its content and facets until a final version is reached. These topics are part of the data on which the experiments are built on.

  Organizers need an interface that allows them to set up the topics to be created, to monitor the creation process, and to publish the topics once they are in the final form.

  Assessors need an interface that allows them to insert and modify the content of a topic, to search the collections of documents to verify that there are relevant documents for the topic, and to discuss together the contents of the topic.

  Note that topics are created by inspecting the documents. The user interface needs to support tasks which reflect the relationships between these two kinds of data;
- *experiment submission*: the participants submit their experiments, which are built using the documents and the topics created in the previous steps. The result of each experiment is a list of retrieved documents in decreasing order to relevance for each topic and represents the output of the execution of the IRS developed by the participant. The experiments are part of the data that are produced during an evaluation campaign.

  Participants need an interface that allows them to upload their experiments into the DLS, to validate them, e.g. to check that the correct document identifiers have been used or that no topic has been skipped, and to provide all the necessary information for describing their experiments.

  Note that experiments are created by starting from documents and topics. The user interface, provides support for checking the correctness of the experiments with respect to topics and documents;
- *creation of pools*: the organizers collect all the experiments submitted by the participants and, using some appropriate sampling technique, select a subset of the retrieved documents to be manually assessed in the next step to determine their actual relevance. The pools are midway between data and information, since they are still quite raw elements but represent a first form of processing of the experiments.

  Organizers need an interface that allows them to select and sample the documents to be inserted in the pool and to dynamically see how the pools change when the selection criteria are modified in order to determine the best strategy for creating the pools.

  This hybrid nature of the pools between data and information is reflected also in the user interface, which explicitly has to show how a pool – i.e. something that relates documents, topics, and experiments – changes when the selection and sampling criteria are modified;
- *relevance assessment*: the organizers and the assessors cooperate together for assessing each document in the pool with respect to the topic, i.e. for determining whether the document is

[9] Topic is the term we adopt for the statements of information needs which are then used by the system to derive their queries; topics can be formulated in various forms according to the particolar tasks for which they will be used.

relevant or not for the given topic. As in the case of the pools, the relevance judgements are midway between data and information, since they are raw elements which constitute an experimental collection but represent human-added information about the relationship between topics and documents of an experiment.

Organizers need an interface that allows them to set up and monitor the relevance assessment process and to publish the relevance judgements once they are in the final form.

Assessors need an interface that allows them to assess the relevance of a document with respect to a topic, to have some basic search functionalities for the documents and topics to assess, and to discuss together in case of topics that may be difficult or ambiguous to assess.

This dual nature of the relevance assessment between data and information is reflected also in the user interface, which explicitly requests the assessors to enter a human judgement (relevant or not relevant) about the relationship between a document and a topic;

- *measures and statistics*: the organizers exploit the relevance assessments in order to compute the performance measures and plots about each experiment submitted by a participant; then, these measurements are used for computing descriptive statistics about the overall behaviour of both an experiment and all the experiments in a given task; furthermore, these measurements are also employed for conducting statistical analyses and tests on the submitted experiments. As discussed above, performance measures are information, since they are the results of a processing on the data; descriptive statistics and hypothesis tests are knowledge, since they provide us with some more insights about the meaning of the obtained performances.

  Organizers need an interface that allows them to perform all the computations and statistical analyses that are need.

  Participants and visitors need an interface that gives access and present performance measurements, plots, descriptive statistics and statistical analyses in a meaningful way in order to facilitate their comprehension and interpretation.
- *scientific production*: both organizers and participants prepare reports where the former describe the overall trends and provide an overview for the evaluation campaign and the latter explain their experiments, the techniques that have been adopted, and the findings. This work usually continues also after the conclusion of the campaign, since the investigation and understanding of the experimental results require deep analysis and reasoning, which usually takes the form of conference papers, journal articles, talks, and discussion among researchers. Furthermore, not only the organizers and the participants but also external visitors may exploit the information resources produced during the evaluation campaign in order to carry out their research activity. As explained above, the outcomes of this process are wisdom.

  Organizers, participants, and visitors need a user interface that provides easy access and meaningful interaction with the information resources, allows them to cite and reference the information resources relevant for their work, and supports the enrichment of the information resources available.

This discussion shows how multifaceted are the needs of the users involved in a large-scale evaluation campaign and how different and complex are the tasks that the DLS used to manage the evaluation campaign has to support. This complexity is also reflected in the user interface, which needs to offer different types of interaction with the system according to the task and user at hand.

The design of a sufficiently functional and responsive user interface must be based on the user needs, analysis of the interaction among users, and the user feedback. Furthermore, a large-scale evaluation campaign involves people from different countries, with different languages and different cultures; this factor has to be taken into account by providing a correct internationalization and localization of the interface in order to lower language and cultural barriers.

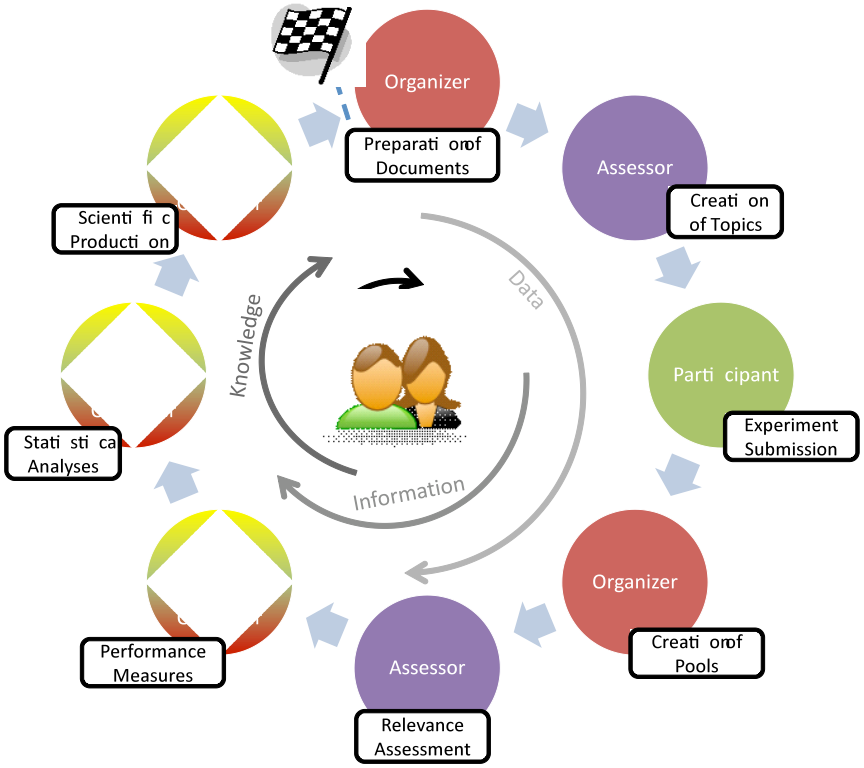

Figure 3: Moving from data to wisdom in an evaluation campaign.

Figure 3 presents the process of moving from data to wisdom in the course of an evaluation campaign, as well as the different actors involved in the various steps. Figure 3 shows that this process is cyclically repeated at each edition of the evaluation campaign; that the role of the different actors is central to this process since their interactions make it possible to pass from one layer to another; and that the different layers are not clearly separated but each step can produce resources that belong to more than one layer.

4 Key Contributions

As observed in the previous section, scientific data, their curation, enrichment, and interpretation are essential components of scientific research. These issues are better faced and framed in the wider context of the curation of scientific data, which plays an important role on the systematic definition of a proper methodology to manage and promote the use of data.

The e-Science Data Curation Report gives the following definition of data curation [30]: “the activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose”. This definition implies that we have to take into consideration the possibility of information enrichment of scientific data, meant as archiving and preserving scientific data so that the experiments, records, and observations will be available for future research, as well as provenance, curation, and citation of scientific data items. The benefits of this approach include the growing involvement of scientists in international research projects and forums and increased interest in comparative research activities. Furthermore, the definition introduced above reflects the importance of some of the many possible reasons for which keeping data is important, for example: re-use of data for new research, including collection based research to generate new science; retention of unique observational data which is impossible to re-create; retention of expensively generated data which is cheaper to maintain than to re-generate; enhancing existing data available for research projects; validating published research results.

As a concrete example in the field of information retrieval, consider the data fusion problem [13], where lists of results produced by different systems have to be merged into a single list. In this context, researchers do not start from scratch, but often experiment their merging algorithms by using the list of results produced in experiments carried out by other researchers. This is the case, for example, of the CLEF 2005 multilingual merging track [14], which provided participants with some of the CLEF 2003 multilingual experiments as list of results to be used as input to their merging algorithms. It is clear that such researchers would benefit by a data curation strategy, which could promote the re-use of existing data and allow data fusion experiments to be traced back to the original list of results and, perhaps, to the analyses and interpretations of them.

However, the Cranfield methodology was developed to create comparable experiments and evaluate the performances of an IRS rather than modeling, managing, and curating the scientific data produced during an evaluation campaign. In the following sections, we discuss some key points that we have taken into consideration when designing the DIRECT digital library system and that extend the current evaluation methodology.

4.1 Conceptual Model

If we consider the definition of experimental collection, it does not take into consideration any kind of conceptual model [37], neither the experimental collection as a whole nor its constituent parts. However, the information space implied by an evaluation campaign needs an appropriate conceptual model that takes into consideration and describes all the entities involved. An appropriate conceptual model is the necessary basis to make the scientific data produced during the evaluation an active part of any information enrichment, such as data provenance and citation. The conceptual model can also be translated into a logical model in order to manage the information of an evaluation campaign by using a robust data management technology. From the conceptual model we can also derive appropriate data formats for exchanging information among organizers and participants.

Figure 4 shows the Unified Modeling Language (UML) schema [27] which represents the conceptual model we have developed and gives an idea of the complexity of the information space involved by an evaluation campaign and for the need of a careful system design. The conceptual model is built around five main areas of modelling:

- *evaluation campaign*: deals with the different aspects of an evaluation forum, such as the evaluation campaigns conducted and the different editions of each campaign, the tracks along which the campaign is organized, the subscription of the participants to the tracks, the topics of each track;
- *collection*: concerns the different collections made available by an evaluation forum; each collection can be organized into various files and each file may contain one or more multimedia documents; the same collection can be used by different tracks and by different editions of the evaluation campaign;
- *experiments*: regards the experiments submitted by the participants and the evaluation metrics computed on those experiments, such as precision and recall;
- *pool/relevance assessment*: is about the pooling method where a set of experiments is pooled and the documents retrieved in those experiments are assessed with respect to the topics of the track the experiments belongs to;
- *statistical analysis*: models the different aspects concerning the statistical analysis of the experimental results, such as the type of statistical test employed, its parameters, the observed test statistic, and so forth.

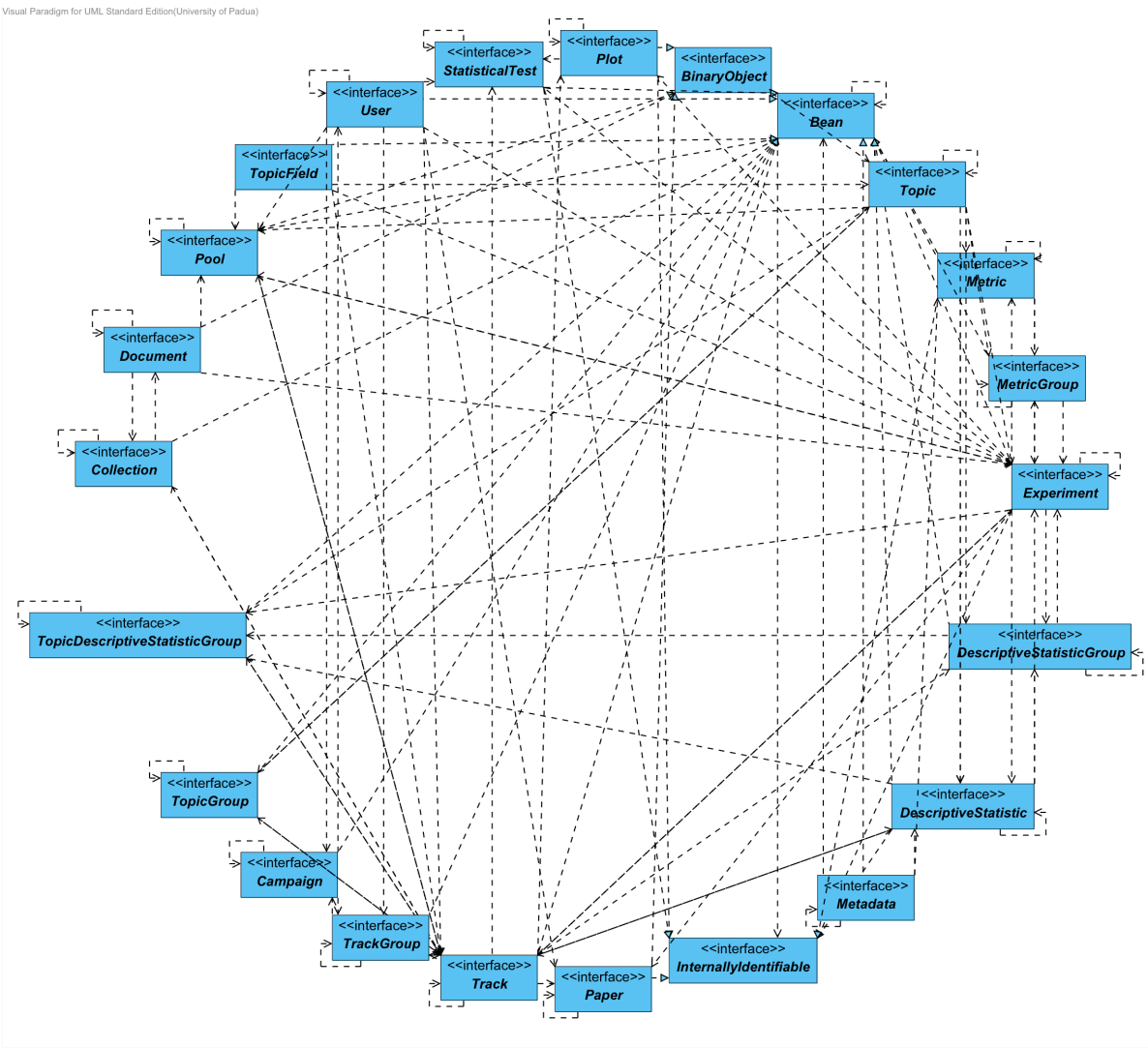

Figure 4: UML conceptual model for the information space of an evaluation campaign.

Each object in the schema has the possibility to be enriched with various metadata objects in order to provide additional information about it; the different metadata objects can comply with different metadata schemes, which can be defined in an easy and extensible way, in order to describe different facets of the annotated object. Moreover, each metadata object can be, in turn, annotated with other metadata objects, so that is possible to have a chain of nested metadata describing a given object.

4.2 Metadata

[7] points out that "metadata descriptions are as important as the data values in providing meaning to the data, and thereby enabling sharing and potential future useful access". Since there is no conceptual model for an experimental collection, also metadata schemes for describing it are lacking. Consider that there are almost no metadata:

- which describe a collection of documents *D*; useful metadata would concern, at least, the creator, the creation date, a description, the context the collection refers to, and how the collection has been created;
- about the topics *T*; useful metadata would regard the creators and the creation date, how the creation process has taken place, if there were any issues, what are the documents the creators have found relevant for a given topic, and so on;

- which describe the relevance judgements *J*; examples of such metadata concern creators and the creation date, what have been the criteria which led the creation of the relevance judgements, what problems have been faced by the assessors when dealing with difficult topics.

The situation is a little bit less problematic when it comes to experiments for which some kind of metadata may be collected, such as which topic fields have been used to create the query, whether the query has been automatically or manually constructed from the topics. TREC also collects more detailed information about the hardware used to run the experiments, what retrieval model has been applied, what algorithms and techniques have been adopted, what kind of stop word removal and/or stemming has been performed, what tunings have been carried out.

A good attempt in this direction is represented by the Reliable Information Access (RIA) Workshop [25], organized by the US 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. 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 [39].

4.3 Unique Identification Mechanism

The lack of a conceptual model causes another relevant consequence: there is no common mechanism for uniquely identify the different digital objects involved in an evaluation campaign, i.e. there is no way to uniquely identify and reference collections of documents, topics, relevance judgements, experiments, and statistical analyses.

The absence of a mechanism to uniquely identify and reference the digital objects of an evaluation campaign prevents us from directly citing that digital object. Indeed, as recognized by [30], the possibility of citing scientific data and their further analysis is an effective way of making scientists and researchers an active part of the digital curation process. This would strengthen the movement from data to wisdom because experimental collections and experiments would become citable and accessible as any other item in the reference list of a paper.

Over the past years, various syntaxes, mechanisms, and systems have been developed to provide unique identifiers for digital objects, among them the following are candidates to be adopted in the unique identification of the different digital objects involved in an evaluation campaign:

- Uniform Resource Identifier (URI) is a compact string of characters for identifying an abstract or physical resource [8] and [9]. The term Uniform Resource Locator (URL) refers to the subset of URIs that identify resources via a representation of their primary access mechanism (e.g., their network “location”), rather than identifying the resource by name or by some other attribute(s) of that resource. The term Uniform Resource Name (URN) refers to the subset of URIs that are required to remain globally unique and persistent even when the resource ceases to exist or becomes unavailable [9];

- Digital Object Identifier (DOI) is a system that provides a mechanism to interoperably identify and exchange intellectual property in the digital environment. DOI conforms to a URI and provides an extensible framework for managing intellectual content based on proven standards of digital object architecture and intellectual property management. Furthermore, it is an open system based on non-proprietary standards [36];
- OpenURL aims at standardizing the construction of "packages of information" and the methods by which they may be transported over networks [32]. Thus, OpenURL is a standard syntax for transporting information (metadata and identifiers) about one or multiple resources within URL, since it provides a syntax for encoding metadata and identifiers, limited to the world of URL [36];
- Persistent URL (PURL)[10]: instead of pointing directly to the location of an Internet resource, a PURL points to an intermediate resolution service that associates the PURL with the actual URL and returns that URL to the client as a standard HyperText Transfer Protocol (HTTP) redirect. The client can then complete the URL transaction in the normal fashion;
- PURL-based Object Identifier (POI)[11] is a simple specification for resource identifiers based on the PURL system and closely related to the use of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) defined by the Open Archives Initiative (OAI)[12] [33]. The POI is a relatively persistent identifier for resources that are described by metadata "items" in OAI-compliant repositories.

An important aspect of all the identification mechanisms described above is that all of them provide facilities for resolving the identifiers. This means that all those mechanisms permit a direct access to each identified digital object starting from its identifier, in this way giving a direct access to an interested researcher to the referenced digital object together with all the information concerning it.

The DOI constitutes a valuable possibility for identifying and referencing digital objects of an evaluation campaign, since there have already been successful attempts to apply it to scientific data and it gives also the possibility of associating metadata to identified digital objects [11] and [34].

Therefore, we adopted the DOI as unique identification mechanism and we register DOIs for different information resources of an evaluation campaign in accordance with mEDRA[13], the multilingual European Registration Agency of DOI. We reserved different DOI prefixes for the following information resources:

- *collections*: prefix 10.2453;
- *topics*: prefix 10.2452;
- *experiments*: prefix 10.2415;
- *pools*: prefix 10.2454;
- *statistical tests*: prefix 10.2455.

The DOI assigned to each of these entities can be resolved to the corresponding information resource which can be easily accessed, as discussed above. Moreover, the DOI allows the citation of the identified information resources, intended as the possibility of explicitly mentioning and making references in the papers to desired experiments, topics, and so on. In this way, it is possible to directly access the data, information, and knowledge which are part of each written production and to better couple them to it.

4.4 Statistical Analyses

[26] points out that, in order to evaluate retrieval performances, we do not need only an experimental collection and measures for quantifying retrieval performances, but also a statistical methodology for judging whether measured differences between retrieval methods can be considered statistically significant.

[10] http://purl.oclc.org/
[11] http://www.ukoln.ac.uk/distributed-systems/poi/
[12] http://www.openarchives.org/
[13] http://www.medra.org/

To address this issue, evaluation campaigns have traditionally supported and carried out statistical analyses, which provide participants with an overview analysis of the submitted experiments. Furthermore, participants may conduct statistical analyses on their own experiments by using either ad-hoc packages, such as IR-STAT-PAK[14], or generally available software tools with statistical analysis capabilities, like R[15], SPSS[16], or MATLAB[17]. However, the choice of performing a statistical analysis or not is left up to each participant who may even not have all the skills and resources needed to perform such analyses. Moreover, when participants perform statistical analyses using their own tools, the comparability among these analyses is not fully guaranteed, in fact, different statistical tests can be employed to analyze the data, or different choices and approximations for the various parameters of the same statistical test can be made.

Therefore, we have provided support and guide to participants to adopt a more uniform way of performing statistical analyses on their own experiments. Indeed, participants cannot only benefit from standard experimental collections which make their experiments comparable, but they can also exploit standard tools for the analysis of the experimental results, which make the analysis and assessment of their experiments comparable too.

As we have already stated, scientific data, their enrichment and interpretation are essential components of scientific research. The Cranfield methodology traces how these scientific data have to be produced, while the statistical analysis of experiments provides the means for further elaborating and interpreting the experimental results. Nevertheless, the current methodology does not require any particular coordination or synchronization between the basic scientific data and the analyses on them, which are treated as almost separate items. On the contrary, researchers could greatly benefit from an integrated vision, where access to a scientific data item could also offer the possibility of retrieving all the analyses and interpretations on it. Furthermore, it should be possible to enrich the basic scientific data in an incremental way, progressively adding further analyses and interpretations.

We carry out statistical analyses concerning the performances of each experiment, such as computing descriptive statistics about the experiment or providing histograms and box plots to analyse the behaviour of the experiment across different topics with respect to different metrics; in this way, we stimulate participants to conduct in-depth analyses of their results by providing them with tools that ease their work. Moreover, we conduct statistical analyses and hypothesis test, such as the Tukey t test, to cross-compare all the experiments submitted for a given task and give the participants the possibility of better understanding their results with respect to the general trends and behaviour for a given task.

5 Architecture of the DIRECT System

As the result of an investigation of user requirements and needs, DIRECT has been designed to meet the following goals:

- 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;

[14] http://users.cs.dal.ca/~jamie/pubs/IRSP-overview.html
[15] http://www.r-project.org/
[16] http://www.spss.com/
[17] http://www.mathworks.com/

- to support internationalization and localization: the application needs to be able to adapt to the language of the user and his country or culturally dependent data, such as dates and currencies.

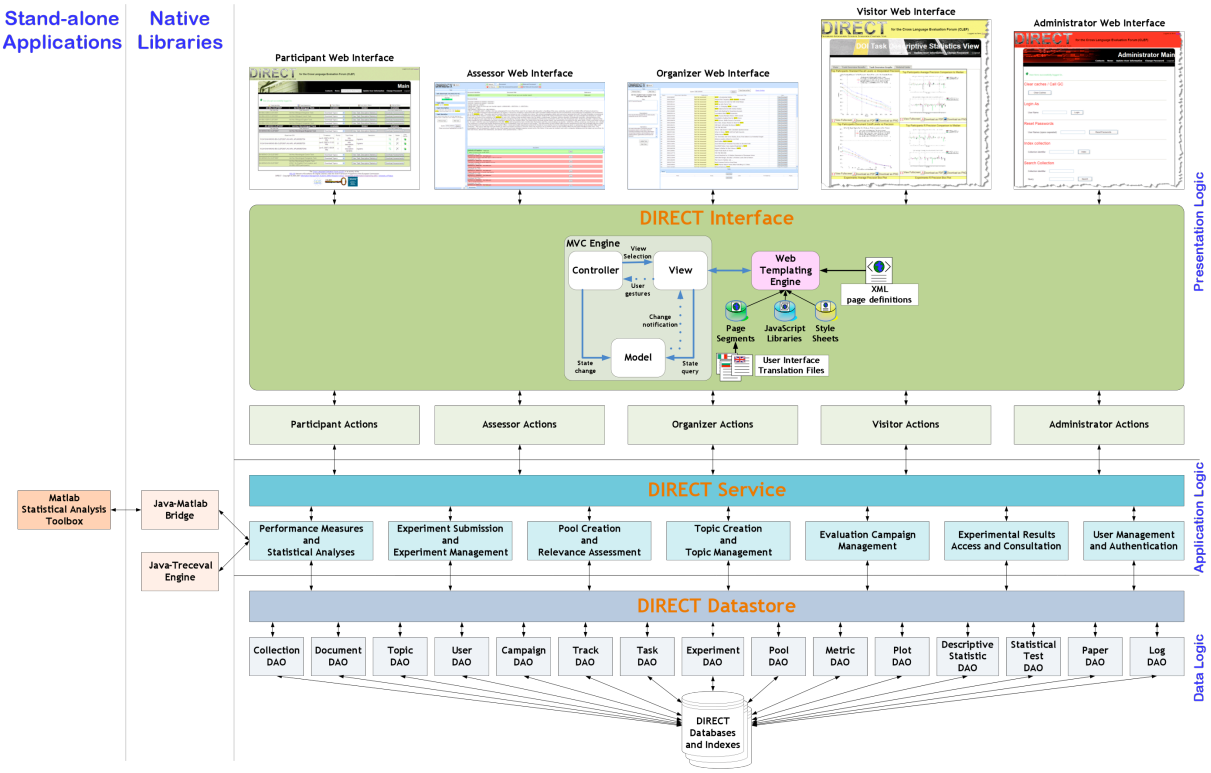

Figure 5: Architecture of the DIRECT system.

Figure 5 shows the architecture of the proposed service. It 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. In this way, the behaviour of the system is designed in a modular and extensible way [3] and [17].

In the following, we briefly describe the architecture shown in Figure 5, from bottom to top.

5.1 Data Logic

The data logic layer 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. We adopt the Data Access Object (DAO) and the Transfer Object (TO)[18] design patterns. The DAO implements the access mechanism required to work with the underlying data source, acting as an adapter between the upper layers and the data source. If the underlying data source implementation changes, this pattern allows the DAO to adapt to different storage schemes without affecting the upper layers.

In addition to the other DAOs, there is the log DAO which fine traces both system and user events. It captures information such as the user name, the Internet Protocol (IP) address of the connecting host, the action that has been invoked by the user, the messages exchanged among the components of the system in order to carry out the requested action, any error condition, and so on. Thus, besides offering us a log of the system and user activities, the log DAO allows us to fine trace the provenance of each piece of data from its entrance in the system to every further processing on it.

[18] http://java.sun.com/blueprints/corej2eepatterns/Patterns/

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.

The data logic layer has been developed by using the Java[19] programming language, which ensures good portability of the system across different platforms. We used the PostgreSQL[20] DataBase Management System (DBMS) for the actual storage of the data and the Lucene[21] library for indexing the document collections and providing the information access capabilities needed for creating topics and performing relevance assessments.

5.2 Application Logic

The application logic 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 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[22] 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. In the case of MATLAB, an additional library is needed to allow our service to access MATLAB in a programmatic way; other applications could require different solutions. A further library provides an interface for our service towards the `trec eval` package[23]. `trec eval` has been firstly developed and adopted by TREC and represents the standard tool for computing the basic performance figures, such as precision and recall.

Finally, the "DIRECT Service" provides the interface logic layer with a 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.

5.3 Interface Logic

The modularity of the components has enormous benefits when building interactive applications, since it helps the designer to better understand and develop each component and modify it without affecting the others. Therefore, we used the Model-View-Controller (MVC) [29] approach as provided by Apache STRUTS[24] framework to clearly separate the following three layers:

- *model layer*: contains the underlying data structures of the application and keeps the state of the application;
- *view layer*: the way the model is presented to the user;
- *controller layer*: manages the interaction between the view and the input devices, such as the keyboard or the mouse, and updates the model accordingly.

Figure 5 shows the architecture of the DIRECT user interface which is a Web-based application in order to be cross-platform and easily deployable and accessible without the need of installing any software on the end-user machines.

The user interface is based on the Java Server Pages (JSP) technology[25]; in addition, we developed a JavaScript[26] library which provides event listeners, manipulation of the Document Object Model

[19] `http://java.sun.com/`
[20] `http://www.postgresql.org/`
[21] `http://lucene.apache.org/`
[22] `http://www.mathworks.com/products/statistics/`
[23] `http://trec.nist.gov/trec_eval/`
[24] `http://struts.apache.org/`
[25] `http://java.sun.com/products/jsp/`

(DOM)[27], and Asynchronous JavaScript Technology and XML (AJAX)[28] support in order to make the user interaction more successful and responsive. In particular, AJAX allows us to make asynchronous calls to the server and to speed up the user interaction by loading only the requested portion of the data without requiring the download of huge amounts of data in one time or the complete refresh of a page when only a part of it has changed.

Moreover, the user interface is made more modular by using the STRUTS TILES[29] templating framework, which allows for a rapid development and reuse of components. As shown in Figure 5, when the browser requests a page, the STRUTS engine asks the TILES engine to put together the page components, according to instructions provided by an eXtensible Markup Language (XML)[30] configuration file. Then, TILES loads the JSP reusable code segments to create the page skeleton, adds the JavaScript libraries needed for enhancing the user interaction, fills the page with the contents provided by the STRUTS controller, applies the necessary Cascading Style Sheets (CSS)[31] for formatting the page, and returns the dynamically created page to the View layer of STRUTS, which, in turn, sends it to the browser.

Finally, we also support the internationalization and localization of the user interface by adapting it to the language and country of the user. As shown in Figure 5, this is implemented by using translation files according to the Java internationalization capabilities[32]. The correct language and country are initially loaded according to the browser settings and, in the case of not supported locales, the interface falls back to a default configuration. The user interface has been translated in the following languages: Bulgarian, Czech, English, Farsi, French, German, Indonesian, Italian, Portuguese, and Spanish.

6 DIRECT: the Running Prototype

DIRECT is successfully adopted in the CLEF campaigns. In the ongoing CLEF 2008 campaign, it is being used by over 130 participants from 20 countries, who have submitted 490 experiments. 80 assessors from over 10 countries have created more than 200 topics in seven different languages and are assessing around 250,000 documents, including documents in languages like Russian, which uses the Cyrillic alphabet, and Farsi, which is written from right to left.

The running prototype of DIRECT is accessible for assessors, participants, organizers, and visitors at the following address: `http://direct.dei.unipd.it/` [18].

In the following section we present some of the functionalities of the DIRECT system and some of its user interfaces.

6.1 Login Page

The login page, shown in Figs 6 to 10, adopts a three columns layout and, besides the login form, provides additional features, like the password recovery form, the automatic detection of user type, and a news ticker. The news, made available through a Really Simple Syndication (RSS) 2.0 feed, informs the user in real time about the latest events in the course of the evaluation campaign; moreover, users can subscribe to the news feed to get notifications and alerts from DIRECT. News is also available in a browser-friendly format, created client-side using eXtensible Stylesheet Language Transformations (XSLT) .

After authentication, the user is automatically redirected to the main page of his user type where a navigation bar allows him to select available tasks, to personalize his account, change email, password,

---

[26] `http://www.ecma-international.org/publications/standards/Ecma-262.htm`
[27] `http://www.w3.org/DOM/`
[28] `http://www.w3.org/TR/XMLHttpRequest/`
[29] `http://struts.apache.org/1.x/struts-tiles/`
[30] `http://www.w3.org/XML/`
[31] `http://www.w3.org/Style/CSS/`
[32] `http://java.sun.com/javase/technologies/core/basic/intl/`

language, and country. Error and information messages from the system are shown in the top of the page with a suitable iconography: green, yellow, or red according to the gravity of the message.

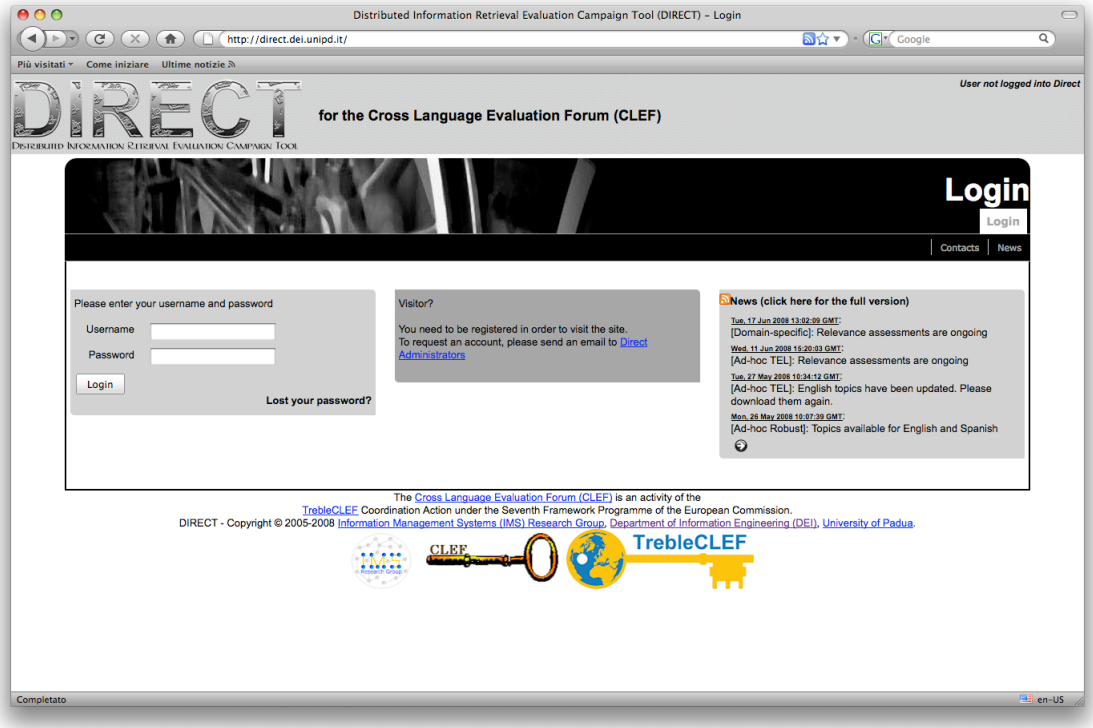

Figure 6: DIRECT login page in English.

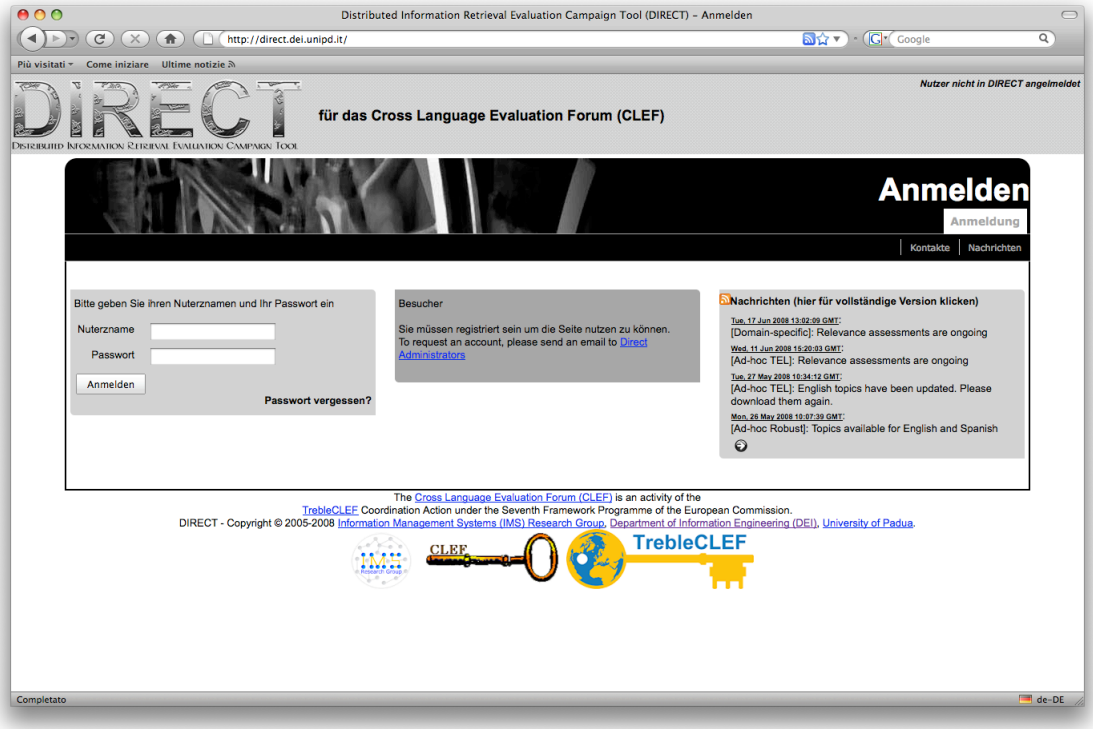

Figure 7: DIRECT login page in German.

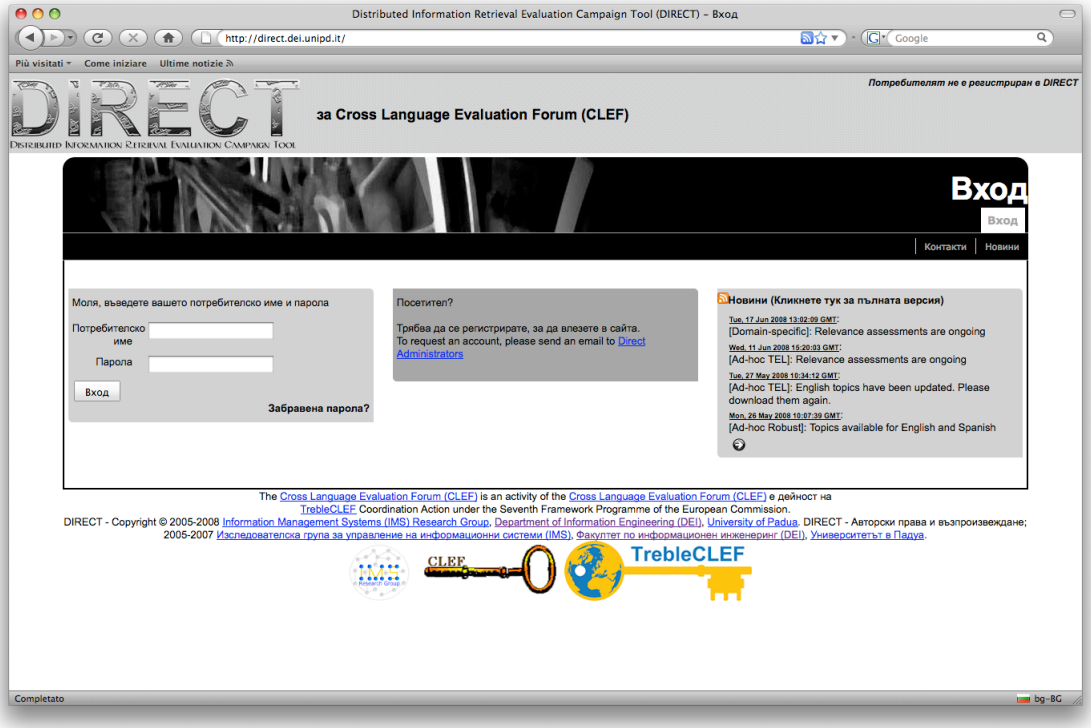

Figure 8: DIRECT login page in Bulgarian.

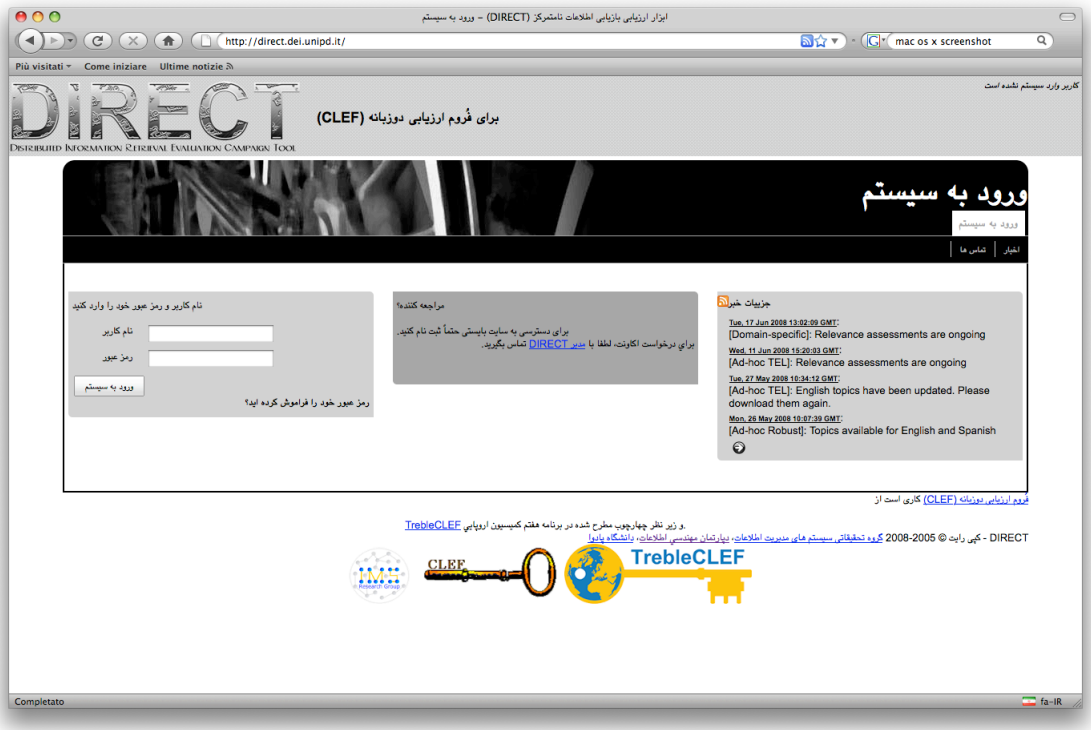

Figure 9: DIRECT login page in Farsi.

6.2 Experiment Management

Figure 10 presents the main page for the experiment management which allows the participant to access all the relevant information about a track, related tasks, topics, and experiments. The interface manages information resources which belong to different levels of the DIKW hierarchy and relates them in a meaningful way. The user can access: the data produced by participants themselves, i.e. the experiments submitted; data produced by assessors, i.e. topics and relevance assessments; information produced by organizers, i.e. performance measures about the experiments submitted by a participant; and, finally, knowledge produced by organizers, i.e. the statistics and statistical analyses about the different tasks of an evaluation campaign.

The interface is based on a set of folding tables, which allow participants to access their experiments, by structuring them in different levels based on a tree structure – tracks, tasks, and experiments – well known to the user. Therefore, the participant can manage his own data by simply selecting and expanding the right level in the tree in order to facilitate the submission, editing, or deletion of an experiment.

Besides experiments, further data are associated with each level of the tree in order to support the participant in accessing additional resources: DIRECT makes available only those topics and relevance assessments that are pertinent for the task currently selected by the participant.

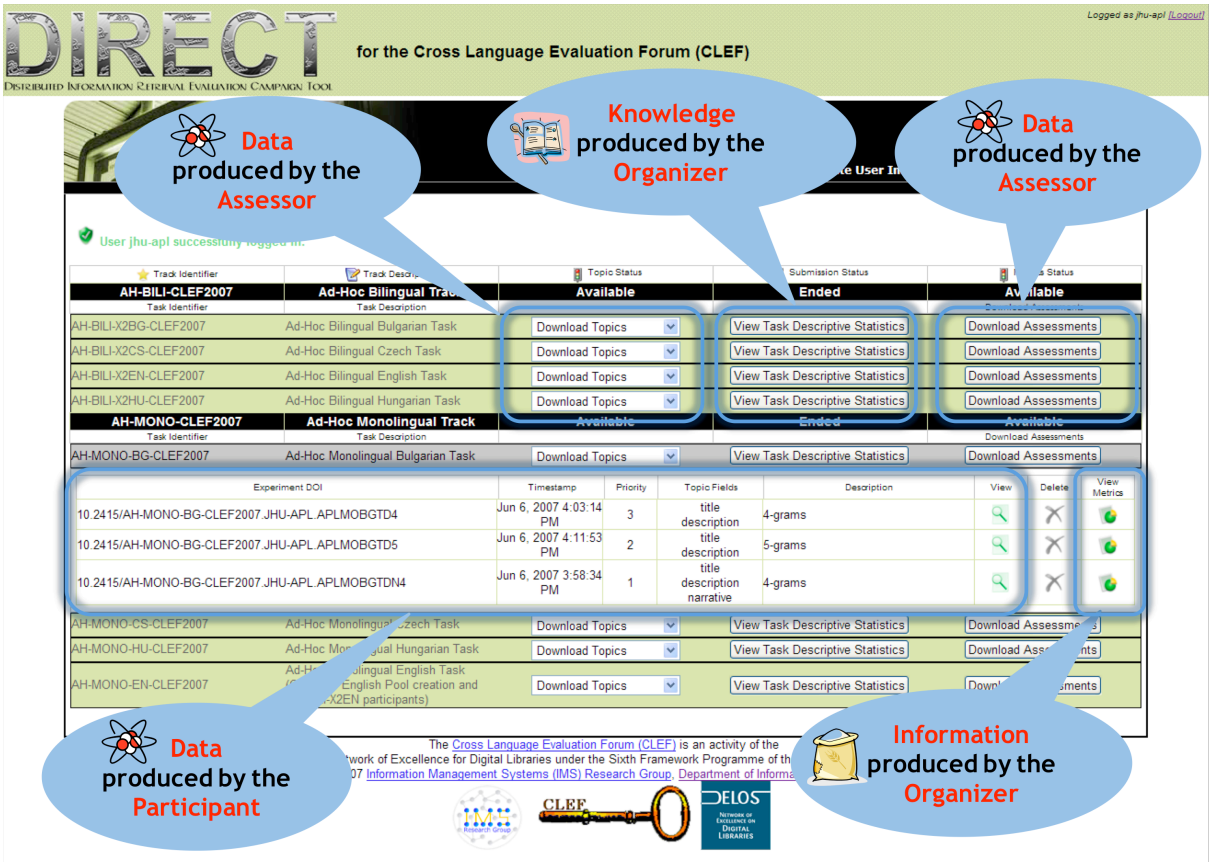

Figure 10: Main page for the experiment management.

Moreover, the system supports the interaction of the participant with the information and knowledge produced by the organizers, i.e. performance measures and statistical analyses, by presenting the appropriate information resources at the correct level in the tree structure.

Finally, following [38] which points out that knowledge is the process through which "individual pieces of data and information (components, concepts) become connected with one another (i.e.

organized) in a network of relations", the system allows users to navigate the interface and access additional information resources, so that they can benefit from this "network of relations".

As an example, Figure 11 shows the information resources offered to the participant when the "View Tasks Descriptive Statistics" button is pressed. In particular, Figure 11 shows some of the plots used to summarize the overall performances achieved in the task and compare the performances of the top participants with respect to the median performances in the task. All these plots can be downloaded and used by participants and visitors, while the numerical data needed to create them can be accessed and downloaded by selecting the "Task Overview Results" tab.

Figure 12 shows the navigation possibilities provided starting from a task: the user can select the topics, experiments, and collections related to the given task. In particular, Figure 12 shows the list of experiments submitted related to the performance measures and plots reported in Figure 11. Thus, the user has the possibility to navigate and browse the experiments submitted by other participants in order to compare them with his own experiments and better understand the pros and cons of the different techniques employed.

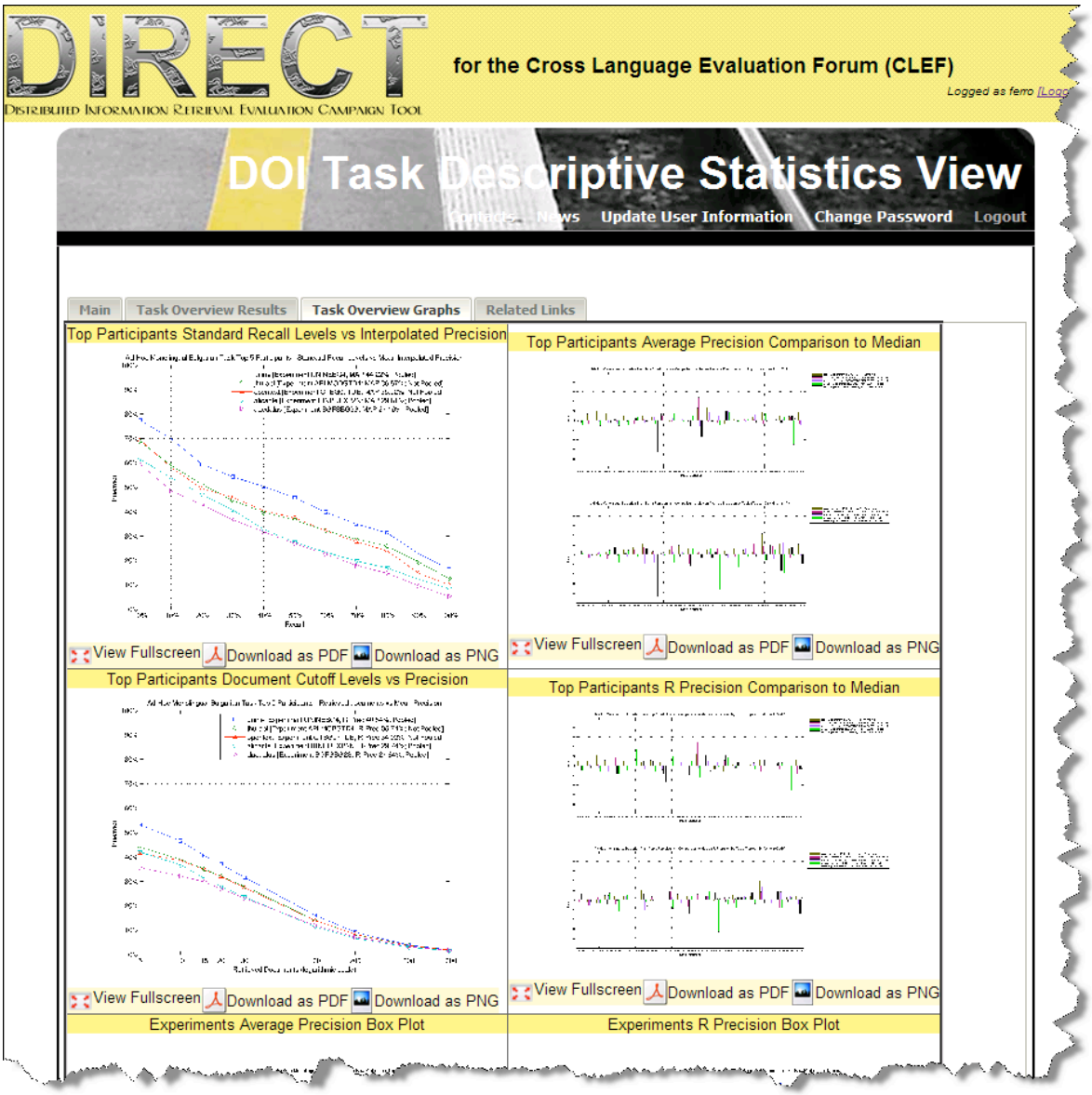

Figure 11: Plots about task overall performances and statistics.

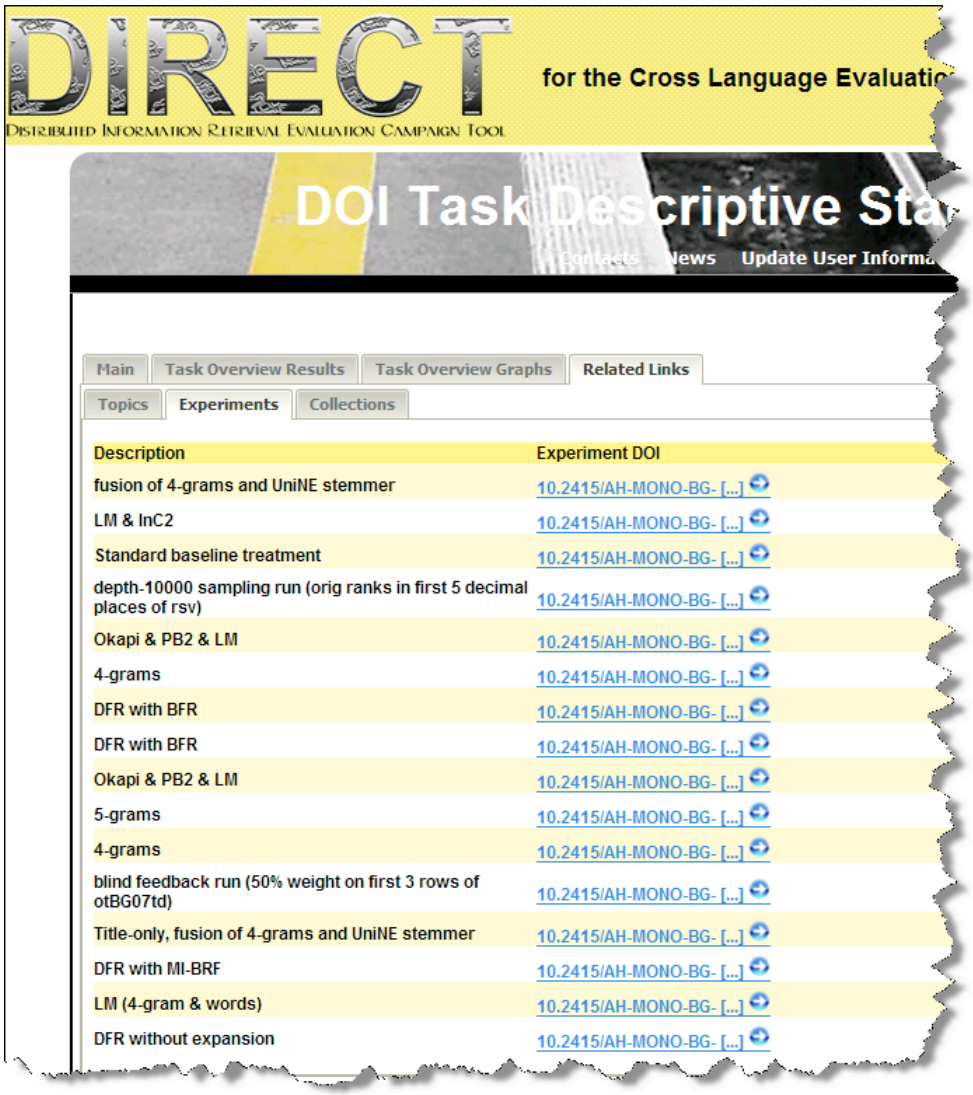

Figure 12: Navigation to other resources: access to the other experiments submitted for a task.

As an additional example, Figure 13 shows the list of topics used in the task related to the performance measures and plots reported in Figure 11. More information about each topic can be obtained by navigating the corresponding link displayed in Figure 13. The results of the navigation are shown in Figure 14, where the contents of a topic are displayed in the different languages selected by the user and accessible for download. Links to the tasks where the topic is used are reported on the right side, so that the user can continue the navigation to other tasks. Therefore, it is, for example, possible to compare the performances achieved for a given topic in different tasks by simply accessing a task, consulting the performances related to it, selecting the topics used in that task, and finally choosing another task that uses the same topics from a list, and then repeating the whole procedure.

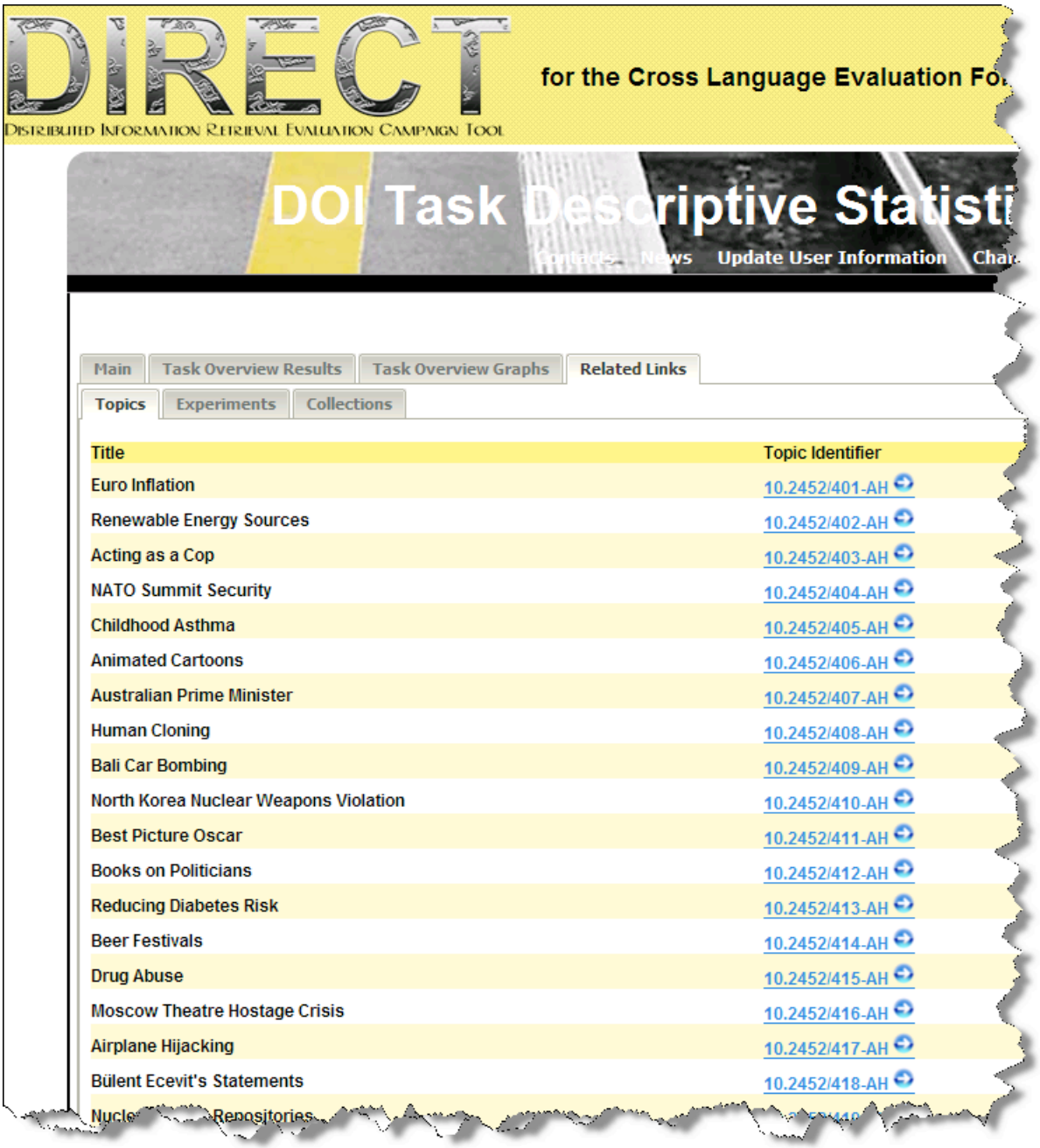

Figure 13: Navigation to other resources: access to the topics related to a task.

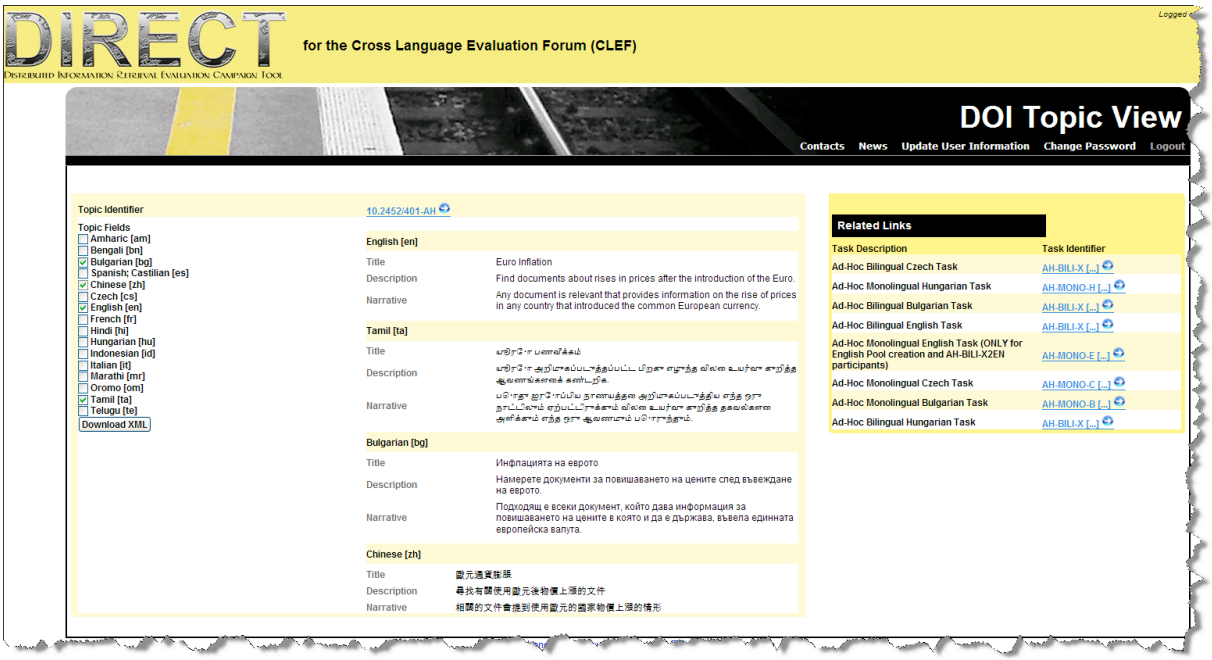

Figure 14: Information about a topic, its different translations, and related tasks.

Note that all the information resources (and more) reported in figures from 10 to 14 are accessible to the participants, assessors and organizers of the CLEF campaigns. Moreover, for those information resources that are not covered by copyright regulations, such as topics and performance measures, DIRECT makes them accessible to the interested visitors who can navigate and consult them by resolving their corresponding DOI. In this way DIRECT contributes to the dissemination of the scientific results obtained within CLEF.

6.3 Topic Creation

Figure 15 shows the main page that the assessor uses for topic creation. The list of topics to be created is presented in table form in order to display the information in a compact and coherent way. For each topic, the topic identifier, a summary, the number of estimated relevant and not relevant documents are shown. Moreover, the table can be folded and expanded so that the user can concentrate on the contents of interest to him without having to read all the data or scroll the whole page. In addition, the system remembers the state of the table by using browser cookies so that the user can find the table opened as he left it, if he reloads the page or logs in again.

Finally, the folding table also helps to increase the loading speed and response times of the interface, as the data needed are only loaded when an assessor ask for a given portion of the table. Asynchronous AJAX calls to the server allow us to retrieve only the portion of information of interest. Since a large amount of data is needed to fill in in this table and it has to be downloaded to the client and rendered by the browser, this choice reduces the amount of data exchanged and, consequently, improves the response times of the interface.

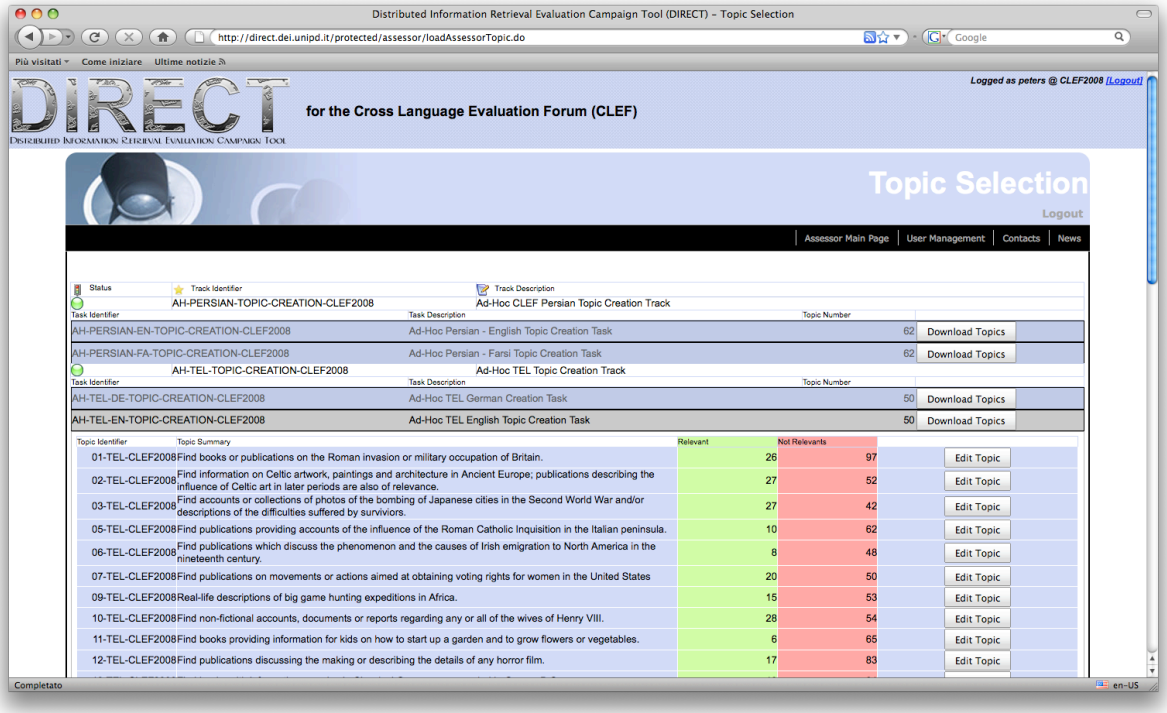

Figure 15: Main page for the creation of new topics.

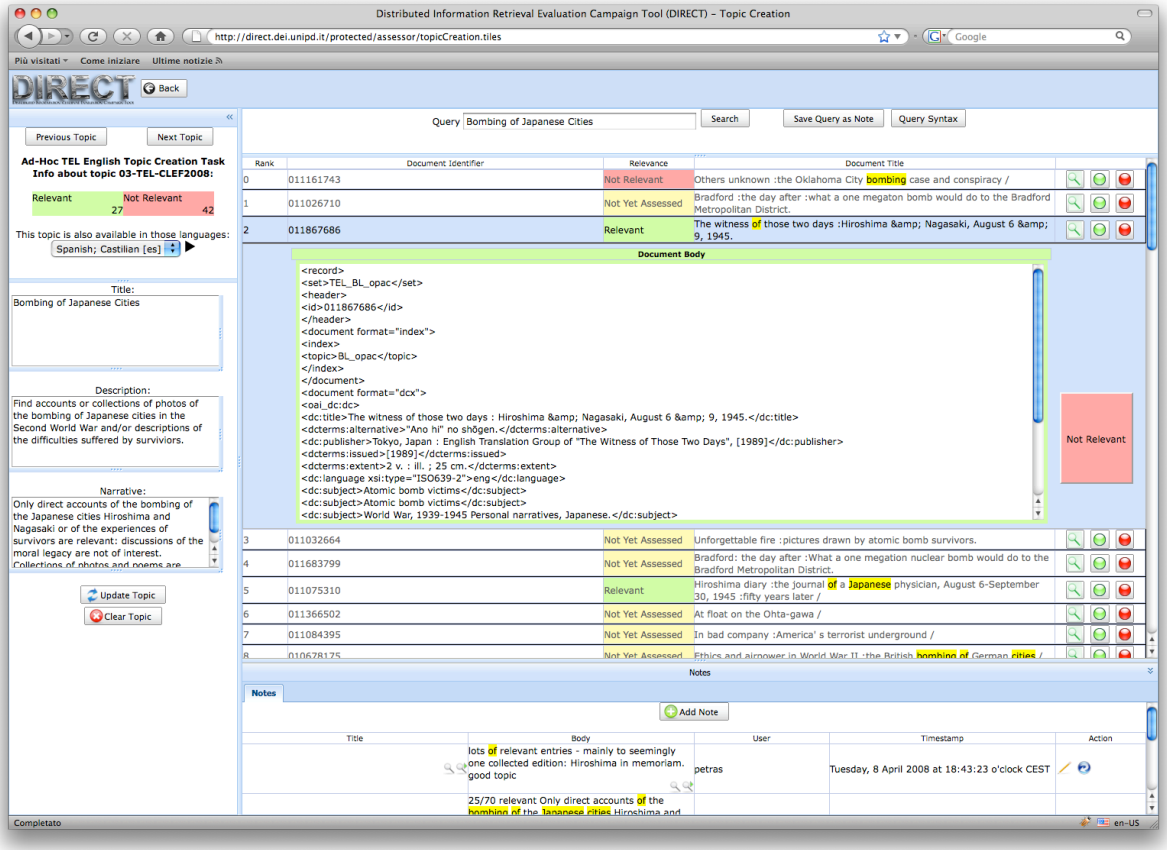

Figure 16: Editing of a new topic.

Once the user clicks the "Edit Topic" button in the main page, the interface for the creation of the topics appears, as shown in Figure 16. On the left side, the assessor can modify the fields of the topic – i.e. title, description, and narrative. At the top of the interface, there is a search box which allows the assessor to search the document collections for the query entered and retrieve a list of documents. This is a fundamental operation, since the assessor has to check that there actually are relevant documents in the collection for the topic under creation.

As can be noted from Figure 16, the list of retrieved documents is shown in the center of the page; for each document, the rank, the document identifier, a relevance assessment, a document summary are shown. The assessor can view the full content of the document by pressing the "View Document" button and can also express a relevance judgment about the document. This judgment is stored in the system and shown whenever the document is retrieved in response to a query. This feature allows assessors to explicitly indicate what document they think might be relevant or not for a given topic. This information is then shared with all the assessors and can contribute to their discussion during the topic creation process. Note that these judgements are different from the ones that are carried out on the pools in the relevance assessment step and which are used for computing the actual performance measures. We could say that the judgements made at topic creation time are a kind of estimate of the relevance of a document but not a final decision.

As shown in the lower part of the window, the assessor has the possibility to add notes and comments about a topic as well as to respond to other assessors' notes. This feature is useful for supporting all the discussion that is usually carried out about a topic during its creation. It is important to remember that topic creation in CLEF is typically a "multilingual" task as, depending on the track, the topics must be designed to be valid (i.e. find a number of relevant documents) in different collections and in different languages. This often implies much discussion between the people responsible for topic creation for the different collections in order to ensure that a given topic is truly "relevant" for all collections involved.

Figure 17 shows that the same interface can be used also for translating the topics from one language to another: the assessor can display the version of the topic available in other languages and enter a new translation of the topic in the "Edit Topic" tab.

6.4 Relevance Assessments

Figure 18 shows the main page that the assessor uses for performing the relevance assessment. The list of topics to be created is presented in table form in order to display the information in a compact and coherent way. For each topic, the topic identifier, a summary, the total number of document to judge, the number of relevant and not relevant documents, and the number of document still to be judged are shown. Moreover, the table can be folded and expanded so that the user can concentrate on the contents of interest to him without having to read all the data or scroll the whole page. In addition, the system remembers the state of the table by using browser cookies so that the user can find the table opened as he left it, if he reloads the page or logs in again.

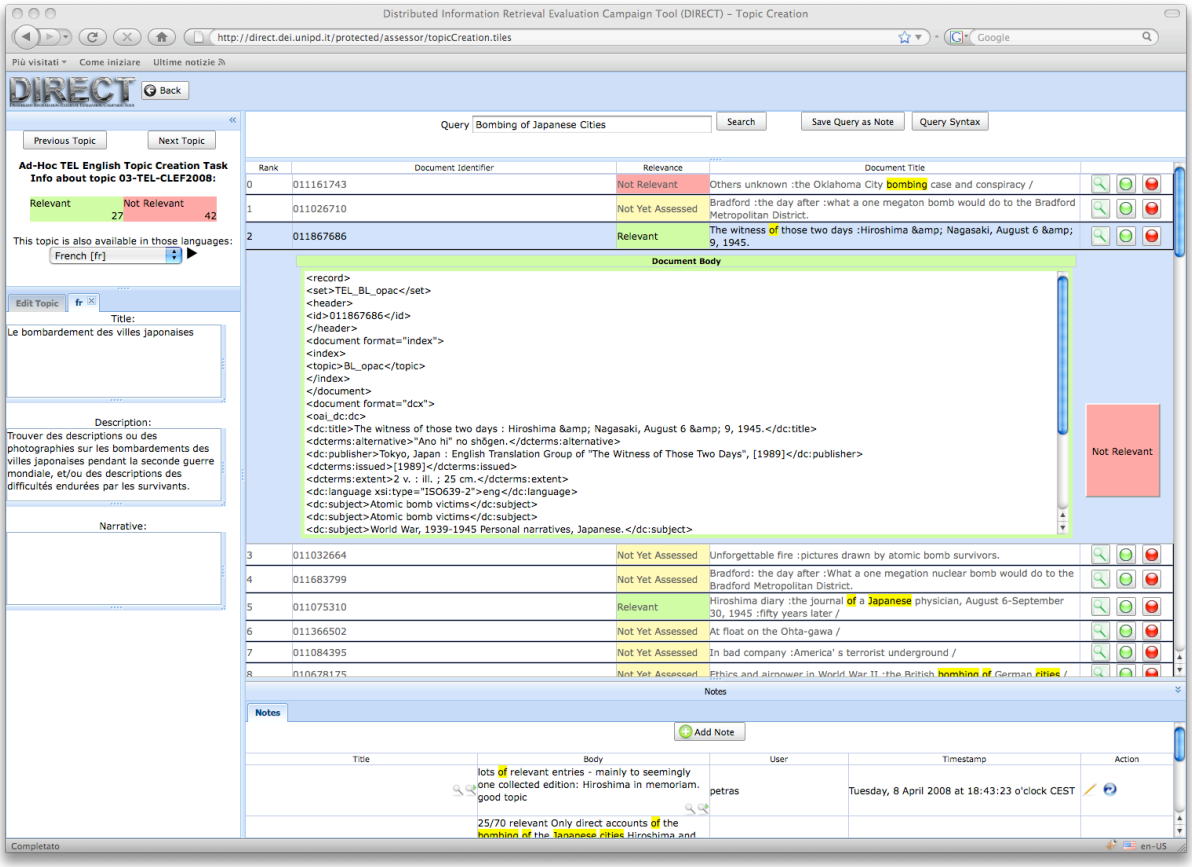

Figure 17: Topic creation: translation from one language to another.

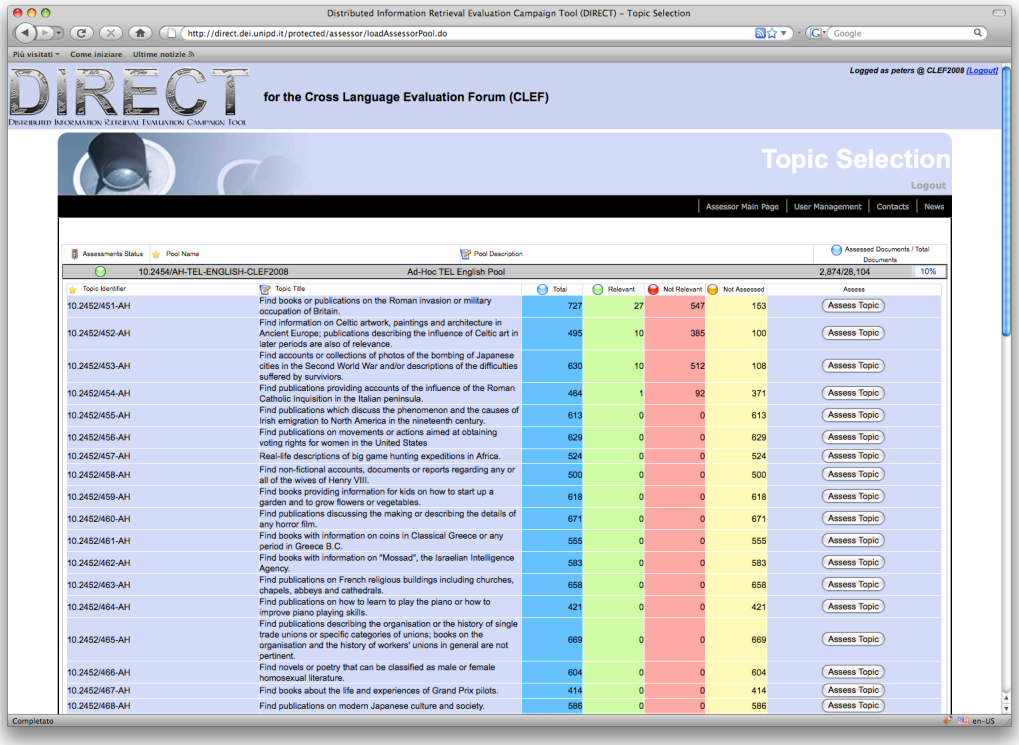

Figure 18: Main page for relevance assessment.

Figure 19 shows the interface for relevance assessments, which is displayed after pressing the "Assess Topic" button. Pools are data, since they are a sampling of the submitted experiments and suggest possible relations among topics and documents in terms of which documents have been retrieved in response to a given topic. On the other hand, relevance judgments are human-added information, since they set the actual relation between a topic and a document, specifying whether a document is relevant or not for a given topic. The outcome of the relevance assessment step is thus the passing from the data contained in the pool to the information contained in the relevance judgments.

The aim of the interface is to support the creation of this information. In the top left corner it is possible to read information about the topic: title, description and narrative are reported, and the status of the assessment task is shown as a progress bar, which gives the percentage of assessments already done. A search form is provided to find terms that occur in both the topic and the selected document. In particular the last submitted query can be saved and automatically repeated on subsequent document selections. Occurrences of the query terms are highlighted in yellow both in the topic and in the document content in order to facilitate the work of the assessor.

Navigation through the documents is facilitated by a set of buttons in the top bar allowing the user to quickly find the next not assessed, relevant or not relevant document.

The selected document is shown at the center of the page, reporting its identifier, title and relevance status. In addition, a highlighting frame flowing up and down over the list of the documents at the bottom of the page shows which document the user is reading in relation to all the documents pooled for that topic.

Specific sets of buttons are also provided at the top of the page to help the user make the assessment task in an intuitive, quick, and useful way. They are characterized by the use of two colours: green to set the relevant status, and red for not relevant. When an assessment is performed, the row at the bottom of the page for the assessed document changes colour accordingly, and the highlighting frame automatically moves to the next document to be assessed.

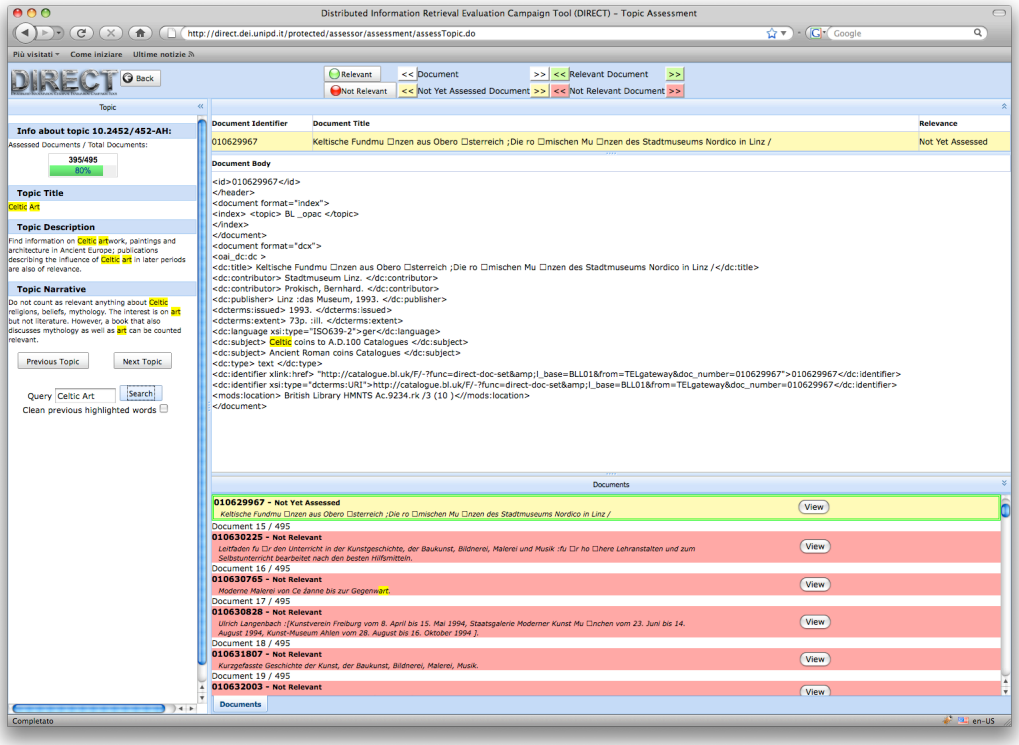

Figure 19: Relevance assessments for a topic.

7 Conclusions

This document has described the methodology currently adopted for experimental evaluation in the IR field, and has shown how we have extended it using the DIRECT system to include an in-depth management, curation, archiving, and enrichment of the scientific data that are produced in the context of large-scale evaluation campaigns.

We described the approach for maintaining in a digital library system the scientific output of an evaluation campaign, in order to ensure long-term preservation of data, and accessibility over time both by humans and automatic systems. The aim is to support services for the creation, interpretation and use of multidisciplinary and multilingual digital content.

We have presented and discussed an innovative software infrastructure to support the course of an evaluation campaign, the running prototype, DIRECT, and its functionalities. DIRECT has been successfully tested and adopted as reference tool during CLEF evaluation campaigns.

Moreover, DIRECT can be exploited as an effective tool to foster knowledge transfer towards not only the IR research community but also relevant application communities and the industry. Indeed, currently, it manages the information resources produced during the CLEF campaigns which are especially interesting for the IR research community. Nevertheless, when used in the context of other TrebleCLEF tasks, such as "Task 3.1 Best practices in system-oriented aspects of MLIA applications" and "Task 4.4 Grid Experiments", DIRECT can manage the experimental results needed to support and explain the indications and guidelines which are the outcomes of these tasks. Indeed, for example, the guidelines could provide hints about what is the best stemmer to be used for a given language and give a reference to the experiment and performance measures, managed by DIRECT, which support the assertion. In this way, system and application developers not only benefit from the findings summarized in the guidelines and best practices but also can have an idea of the actual performances they can expect adopting a solution or another by accessing the only the information relevant to them through DIRECT.

Acknowledgements

We would like to warmly thank all the people who contributed to the translation of the user interface: Petya Osenova and Kiril Simov for Bulgarian; Pavel Pecina for Czech; Abolfazl AleAhmad for Farsi; Jacques Savoy for French; Thomas Mandl for German; Mirna Adriani for Indonesian; Diana Santos and Paulo Rocha for Portuguese; and Julio Villena Romàn for Spanish.

References

[1] Abiteboul, S., Agrawal, R., Bernstein, P., Carey, M., Ceri, S., Croft, B., et al. (2005). The Lowell Database Research Self-Assessment. *Communications of the ACM (CACM), 48* (5), 111-118.

[2] Ackoff, R. L. (1989). From Data to Wisdom. *Journal of Applied Systems Analysis , 16*, 3-9.

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

[4] Agosti, M., Di Nunzio, G. M., & Ferro, N. Scientific Data of an Evaluation Campaign: Do We Properly Deal With Them? In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), *Evaluation of Multilingual and Multi-modal Information Retrieval : Seventh Workshop of the Cross-Language Evaluation Forum (CLEF 2006). Revised Selected Papers*, pp. 11-20. Lecture Notes in Computer Science (LNCS) 4730, Springer, Heidelberg, Germany.

[5] Agosti, M., Di Nunzio, G. M., & Ferro, N. (2007). The Importance of Scientific Data Curation for Evaluation Campaigns. In C. Thanos, F. Borri, & L. Candela (Eds.), *Digital Libraries:*

Research and Development. First International DELOS Conference. Revised Selected Papers, pp. 157-166. Lecture Notes in Computer Science (LNCS) 4877, Springer, Heidelberg, Germany.

[6] Agosti, M., Di Nunzio, G. M., Ferro, N., Harman, D., & Peters, C. (2007). The Future of Large-scale Evaluation Campaigns for Information Retrieval in Europe. In N. Fuhr, L. Kovacs, & C. Meghini (Eds.), *Proc. 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2007)*, pp. 509-512. Lecture Notes in Computer Science (LNCS) 4675, Springer, Heidelberg, Germany.

[7] Anderson, W. L. (2004). Some Challenges and Issues in Managing, and Preserving Access To, Long-Lived Collections of Digital Scientific and Technical Data. *Data Science Journal , 3*, 191-202.

[8] Berners-Lee, T. (1994). Universal Resource Identifiers in WWW. RFC 1630.

[9] Berners-Lee, T., Fielding, R., Irvine, U. C., & Masinter, L. (1998). Uniform Resource Identifiers (URI): Generic Syntax. RFC 2396.

[10] Braschler, M., Di Nunzio, G. M., Ferro, N., Gonzalo, J., Peters, C., & Sanderson, M. (2007). From CLEF to TrebleCLEF: promoting Technology Transfer for Multilingual Information Retrieval. In C. Thanos, F. Borri, & A. Launaro (Eds.), *Second DELOS Conference - Working Notes*. ISTI-CNR, Gruppo ALI, Pisa, Italy.

[11] Brase, J. (2004). Using Digital Library Techniques - Registration of Scientific Primary Data. In R. Heery, & L. Lyon (Eds.), *Proc. 8th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2004)*, pp. 488-494. Lecture Notes in Computer Science (LNCS) 3232, Springer, Heidelberg, Germany.

[12] Cleverdon, C. W. (1997). The Cranfield Tests on Index Languages Device. In K. Spärck Jones, & P. Willet (Eds.), *Readings in Information Retrieval*, pp. 47-60. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA.

[13] Croft, W. B. (2000). Combining Approaches to Information Retrieval. In W. B. Croft (Ed.), *Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval*, pp. 1-36. Kluwer Academic Publishers, Norwell (MA), USA.

[14] Di Nunzio, G. M. and Ferro, N. (2005). DIRECT: a System for Evaluating Information Access Components of Digital Libraries. In A. Rauber, S. Christodoulakis, and A. Min Tjoa (Eds.), *Proc. 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)*, pp. 483–484. Lecture Notes in Computer Science (LNCS) 3652, Springer, Heidelberg, Germany.

[15] Di Nunzio, G. M. and Ferro, N. (2006). Scientific Evaluation of a DLMS: a service for evaluating information access components. In J. Gonzalo, C. Thanos, M. F. Verdejo, and R. C. Carrasco (Eds.), *Proc. 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006)*, pp. 536–539. Lecture Notes in Computer Science (LNCS) 4172, Springer, Heidelberg, Germany.

[16] Di Nunzio, G. M., Ferro, N., Jones, G. J., & Peters, C. (2006). CLEF 2005: Ad Hoc Track Overview. In C. Peters, F. C. Gey, J. Gonzalo, G. J. Jones, M. Kluck, B. Magnini, et al. (Eds.), *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross--Language Evaluation Forum (CLEF 2005). Revised Selected Papers*, pp. 11-36. Lecture Notes in Computer Science (LNCS) 4022, Springer, Heidelberg, Germany.

[17] Dussin, M., & Ferro, N. (2008). Design of a Digital Library System for Large-Scale Evaluation Campaigns. In B. Christensen-Dalsgaard, D. Castelli, J. K. Lippincott, B. Ammitzbøll Jurik (Eds.), *Proc. 12th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2008)*. Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany.

[18] Dussin, M., & Ferro, N. (2008). DIRECT: Applying the DIKW Hierarchy to Large-Scale Evaluation Campaigns. In R. Larsen, A. Paepcke, J. L. Borbinha, and M. Naaman (Eds.), *Proc.*

8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008), p. 424. ACM Press, New York, USA.

[19] Dussin, M., & Ferro, N. (2008). The Design of the User Interface of a Scientific DLS in the Context of the Data, Information, Knowledge, and Wisdom Hierarchy. In M. Agosti, F. Esposito, and C. Thanos (Eds.), *Post-proceedings of the Fourth Italian Research Conference on Digital Library Systems (IRCDL 2008)*, pp. 105-113. ISTI-CNR at Gruppo ALI, Pisa, Italy.

[20] European Commission Information Society and Media. (2006). *i2010: Digital Libraries.* `http://europa.eu.int/information_society/activities/digital_libraries/doc/brochures/dl_brochure_2006.pdf`.

[21] Fuhr, N., Hansen, P., Micsik, A., & Sølvberg, I. Digital Libraries: A Generic Classification Scheme. In P. Constantopoulos, & I. T. Sølvberg (Eds.), *Proc. 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001)*, pp. 187-199. Lecture Notes in Computer Science (LNCS) 2163, Springer, Heidelberg, Germany.

[22] Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., et al. (2007). Evaluation of Digital Libraries. *International Journal on Digital Libraries , 8* (1), 21-38.

[23] Gradmann, S. (2007). Interoperability of Digital Libraries: Report on the work of the EC working group on DL interoperability. In *Seminar on Disclosure and Preservation: Fostering European Culture in The Digital Landscape.* {National Library of Portugal, Directorate-General of the Portuguese Archives, Lisbon, Portugal, `http://bnd.bn.pt/seminario-conhecer-preservar/doc/Stefan%20Gradmann.pdf`.

[24] Harman, D. K., & Voorhess, E. M. (Eds.). (2005). *TREC. Experiment and Evaluation in Information Retrieval.* MIT Press, Cambridge (MA), USA.

[25] Harman, D., & Buckley, C. (2004). SIGIR 2004 Workshop: RIA and "Where can IR go from here?". *ACM SIGIR Forum , 38* (2), 45-49.

[26] Hull, D. A. (1993). Using Statistical Testing in the Evaluation of Retrieval Experiments. In R. Korfhage, E. Rasmussen, & P. Willett (Eds.), *Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993)*, pp. 329-338. ACM Press, New York, USA.

[27] International Organization for Standardization (2005). *Information technology - Open Distributed Processing - Unified Modeling Language (UML) Version 1.4.22.* Recommendation ISO/IEC 19501:2005.

[28] Ioannidis, Y., Maier, D., Abiteboul, S., Buneman, P., Davidson, S., Fox, E. A., et al. (2005). Digital library information-technology infrastructures. *International Journal on Digital Libraries , 5* (4), 266-274.

[29] Krasner, G. E., & Pope, S. T. (1988). A Cookbook for Using the Model-View-Controller User Interface Paradigm in Smalltalk-80. *Journal of Object-Oriented Programming , 1* (3), 26-49.

[30] Lord, P., & Macdonald, A. (2003). *e-Science Curation Report. Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision.* The JISC Committee for the Support of Research (JCSR). `http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf`.

[31] National Science Board (2005). *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century (NSB-05-40).* National Science Foundation (NSF), `http://www.nsf.gov/pubs/2005/nsb0540/`.

[32] National Information Standards Organization (2005). *ANSI/NISO Z39.88 - 2004 - The OpenURL Framework for Context-Sensitive Services.* NISO, `http://www.niso.org/standards/standard_detail.cfm?std_id=783`.

[33] Open Archives Initiative (2004). *The Open Archives Initiative Protocol for Metadata Harvesting - Version 2.0.* (C. Lagoze, H. Van De Sompel, M. Nelson, & S. Warner, Eds.) http://www.openarchives.org/OAI/openarchivesprotocol.html.

[34] Paskin, N. (2005). Digital Object Identifiers for Scientific Data. *Data Science Journal , 4*, 12-20.

[35] Paskin, N. (Ed.). (2006). *The DOI Handbook - Edition 4.4.1.* International DOI Foundation (IDF). http://dx.doi.org/10.1000/186.

[36] Science, W. G. (2006). *FROM DATA TO WISDOM: Pathways to Successful Data Management for Australian Science.* Report to Ministers Science, Engineering and Innovation Council (PMSEIC), http://www.dest.gov.au/sectors/science_innovation/publications_resources/profiles/Presentation_ Data_for_Science.htm.

[37] Tsichritzis, D. C., & Lochovsky, F. H. (1982). *Data Models.* Prentice Hall, Englewood Cliffs (N.J), USA.

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

[39] Zobel, J. (1998). How Reliable are the Results of Large-Scale Information Retrieval Experiments. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)* (pp. 307-314). ACM Press, New York, USA.