A Proposal to Extend and Enrich the Scientific Data Curation of Evaluation Campaigns

Maristella Agosti   Giorgio Maria Di Nunzio   Nicola Ferro
Department of Information Engineering – University of Padua
Via Gradegnigo 6/a, 35131 Padova, Italy
{agosti, dinunzio, ferro}@dei.unipd.it

Abstract

This paper examines the current way of keeping the data produced during an evaluation campaign of Information Retrieval Systems (IRSs) and highlights some shortenings of it. In particular, the Cranfield methodology has been designed for creating comparable experiments and evaluating the performances of IRS rather than modeling and managing the scientific data *produced during an evaluation campaign.*

The data produced during an evaluation campaign of IRSs are valuable scientific data, and as a consequence, their lineage *should be tracked since it allows us to judge the quality and applicability of information for a given use; those data should be* enriched *progressively adding further analyses and interpretations on them; it should be possibile to* cite *them and their further elaboration, since this is an effective way for explicitly mentioning and making references to useful information, for improving the cooperation among researchers and to facilitate the transfer of scientific and innovative results from research groups to the industrial sector.*

Keywords: *Experimentation, Scientific Data, Data Curation, Long-term Preservation*

1 Introduction

This paper addresses the experimental evaluation approach adopted by the *Information Retrieval (IR)* research field in the light of the challenges posed by the increasing attention for the management, preservation and access to scientific data. We describe how this increasing attention impacts both the IR evaluation methodology and the way in which the data of the evaluation campaigns are organized and maintained over time. And, we explain the concrete steps we have undertaken in order to contribute to an organic and systematic extension of the current IR evaluation methodology, since we have designed a conceptual model and developed an effective architecture for an innovative software infrastructure to support the course of an evaluation campaign. To reach its aim and to present the corresponding findings, the paper is organized as follows: Section 2 introduces the motivations and the objectives of our research work; Section 3 discusses possible ways of extending the current evaluation methodology; Section 4 describes the conceptual model of the information space involved by and a software infrastructure for an evaluation campaign; Section 5 provides information about the running system; finally, Section 6 draws some conclusions.

2 The IR Experimental Evaluation

2.1 Methodological Viewpoint

The current approach for laboratory evaluation of information access systems relies on the Cranfield methodology, which makes use of *experimental collections* [11]. An experimental collection is a triple $\mathcal{C} = (D, T, J)$, 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 $\mathcal{C}$ 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

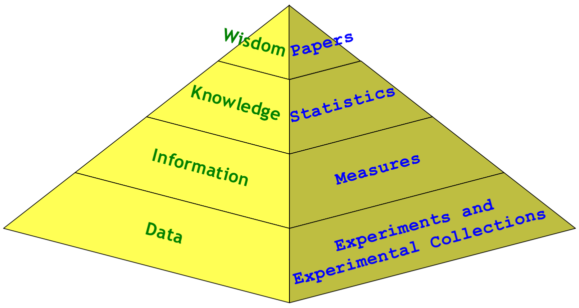


Figure 1. DIKW hierarchy with respect to IR experimental evaluation.

precision and recall, that 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.

We can frame the above mentioned scientific data in the context of the *Data, Information, Knowledge, Wisdom (DIKW)* hierarchy [2, 34], represented in figure 1:

- *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 to 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 [34], "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 to the case of IR experimental evaluation. Indeed, experiments (data) and performance measurements (information) are usually generated in an automatic way by IRSs, 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 for the basic question whether the Cranfield methodology is able to support an experimental approach where the whole process from data to wisdom is taken into account.

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, [22] 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 [3, 4]. Finally, [25] 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.

2.2 Infrastructural Viewpoint

When it comes to infrastructural aspects of such evaluation methodology, the experimental evaluation is usually carried out in important international evaluation campaigns which bring research groups together, provide them with the means for measuring the performances of their systems, discuss and compare their results. *Text REtrieval Conference (TREC)*[1] has been the first initiative in this field and has laid the groundwork for the other subsequent initiatives; TREC developed a common evaluation procedure in order to compare IRSs by measuring the effectiveness of different techniques, and to discuss how differences between systems affected performances [20]. After TREC, other international important initiatives have been launched, in particular *Cross-Language Evaluation Forum (CLEF)* and *NII-NACSIS Test Collection for IR Systems (NTCIR)*. CLEF[2] aims at evaluating *Cross Language Information Retrieval (CLIR)* systems that operate on European languages in both monolingual and multilingual contexts. NTCIR[3] is the Asian counterpart of CLEF where the traditional Chinese, Korean, Japanese, and English languages are the basis of the evaluation of cross-lingual tasks.

The growing interest in the proper management of scientific data has been brought to general attention by different world organizations, among them the European Commission, the US National Scientific Board, and the Australian Working Group on Data for Science. The EC 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” [18]. 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” [25]. 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” [33].

The above mentioned observations suggest that considering the IR experimental evaluation as a source of scientific data requests not only to re-think about the evaluation methodology itself but also to re-consider the way in which this methodology is carried out and in which evaluation campaigns are organized. Indeed, changes to the IR evaluation methodology need to be properly supported by an organizational, hardware, and software infrastructures which allow for management, search, access, curation, enrichment, and citation of the produced scientific data.

This change involves also the organizations which set up the evaluation campaigns, since they have not only to provide such infrastructure but also to participate in the design and development of it. In fact, as highlighted by [25], they should take a leadership 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 provide them the means for discussing and comparing their work, an evaluation campaign should take care also of defining guiding principles, policies, best practices for making use of the scientific data produced during the evaluation campaign itself.

3 Extending Evaluation

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 [24]: “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

[1] `http://trec.nist.gov/`

[2] `http://clef.isti.cnr.it/`

[3] `http://research.nii.ac.jp/ntcir/index-en.html`

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, please consider the data fusion problem [12], 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 they 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 [17], 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 researchers of this field 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 about them.

On the other hand, the Cranfield methodology was developed to create comparable experiments and evaluating 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 we propose to explicitly take into consideration for extending the current evaluation methodology.

3.1 Conceptual Model and Metadata

If we consider the definition of experimental collection, it does not take into consideration any kind of conceptual model [31] of neither the experimental collection as a whole nor its constituent parts. Whereas, the information space implied by an evaluation campaign needs an appropriate conceptual model which takes into consideration and describes all the entities involved by the evaluation campaign. In fact, an appropriate conceptual model is the necessary basis to make the scientific data produced during the evaluation an active part of all those information enrichments, as data provenance and citation. The conceptual model can be also translated into an appropriate logical model in order to manage the information of an evaluation campaign by using a robust data management technology. Finally, from this conceptual model we can derive also appropriate data formats for exchanging information among organizers and participants.

Moreover, [5] 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 [15];
- 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 [15].

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 and, in some tracks of TREC, some information about the hardware used to run the experiments. Nevertheless, a better description of the experiments could be achieved if we take into consideration 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 [10, 19], 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 have been performed and valuable information about them have been 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 which are not explicitly modeled 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 have been used to compute a performance measure, which relevance judgements have been used to compute a performance measure, and when the performance measure has been computed.

Similar metadata could be useful also for descriptive statistics and statistical analyses.

All these 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 [35].

3.2 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 to collections of documents, topics, relevance judgements, experiments, and statistical analyses.

The absence of a mechanism for uniquely identify and reference the digital objects of an evaluation campaign prevent us from directly citing those digital object. Indeed, as recognized by [24], the possibility of citing scientific data and their further elaboration is an effective way for making scientists and researchers an active part of the digital curation process. Moreover, this opportunity would strengthen the passing from data to wisdom, discussed in Section 2, 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 [6, 7]. 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 [7];

- *Digital Object Identifier (DOI)* is a system which 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 [29];

- OpenURL aims at standardizing the construction of “packages of information” and the methods by which they may be transported over networks [26]. Thus, OpenURL is a standard syntax for transporting information (metadata and identifiers) about one or multiple resources within URLs, since it provides a syntax for encoding metadata and identifiers, limited to the world of URLs [29];

- *Persistent URL (PURL)*[4]: 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)*[5] 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)*[6] [27]. 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 [9, 28].

[4] `http://purl.oclc.org/`

[5] `http://www.ukoln.ac.uk/distributed-systems/poi/`

[6] `http://www.openarchives.org/`

3.3 Statistical Analyses

[21] 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.

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; recently results of this kind have been presented in [17, 23, 32]. Furthermore, participants may conduct statistical analyses on their own experiments by using either ad-hoc packages, such as IR-STAT-PAK[7], or generally available software tools with statistical analysis capabilities, like R[8], SPSS[9], or MATLAB[10]. However, the choice of whether performing a statical 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 could not be fully granted, 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.

In developing an infrastructure for improving the support given to participants by an evaluation campaign, it could be advisable to add some form of support and guide to participants for adopting a more uniform way of performing statistical analyses on their own experiments. If this support is added, participants can not 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 would make the analysis and assessment of their experiments comparable too.

As recalled in Section 2, scientific data, their enrichment and interpretation are essential components of scientific research. The Cranfield methodology traces out how these scientific data have to be produced, while the statistical analysis of experiments provide the means for further elaborating and interpreting the experimental results. Nevertheless, the current methodologies does not require any particular coordination or synchronization between the basic scientific data and the analyses on them, which are treated as almost separated items. On the contrary, researchers would greatly benefit from an integrated vision of them, where the 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 on them.

4 Evaluation Campaign Infrastructure

4.1 Conceptual Model

As discussed in the previous section, we need to design and develop a proper conceptual model of the information space involved by an evaluation campaign. Indeed, this conceptual model provide us with the basis needed to offer all the information enrichment and interpretation features described above.

Figure 2 shows the *Unified Modeling Language (UML)* schema [30, 8] which represents the conceptual model we have developed and gives and 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 conducted evaluation campaigns 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.

Each object in the schema has the possibility to be enriched with various metadata objects in order to

[7] `http://users.cs.dal.ca/~jamie/pubs/IRSP-overview.html`

[8] `http://www.r-project.org/`

[9] `http://www.spss.com/`

[10] `http://www.mathworks.com/`

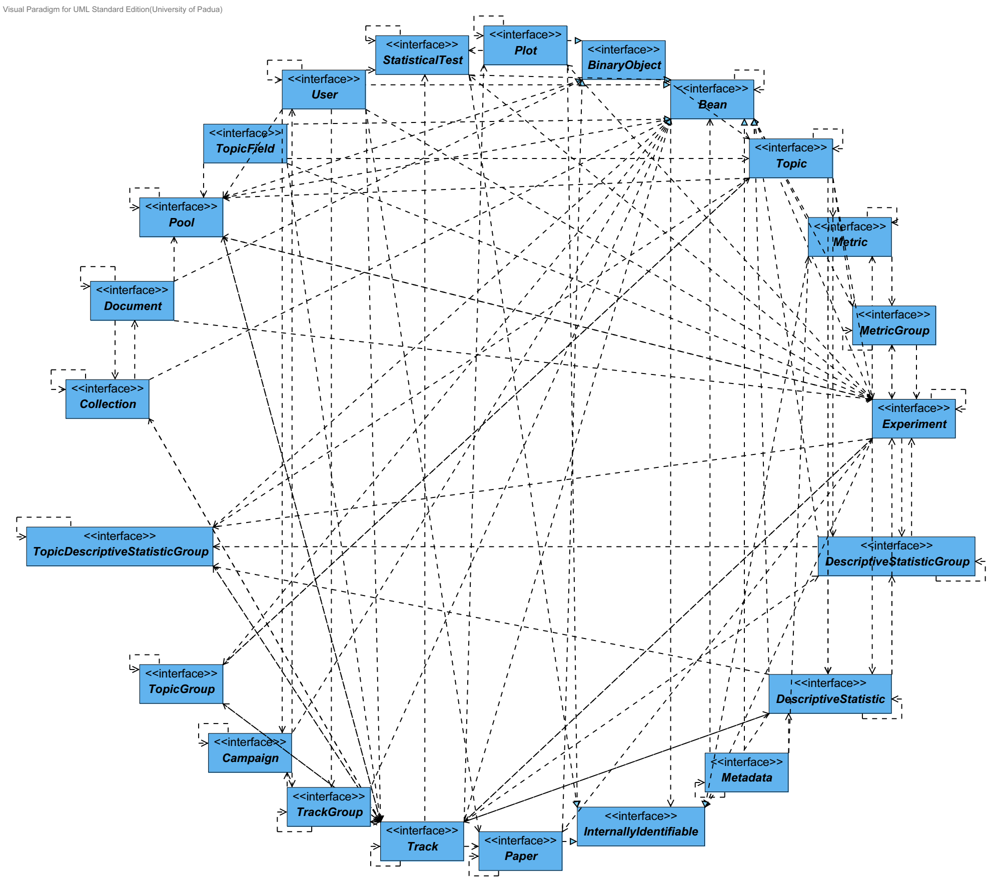


Figure 2. UML conceptual model for the information space of an evaluation campaign.

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 Architecture

Figure 3 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 behavior of the service by isolating specific functionalities at the proper layer. In this way, the behavior of the system is designed in a modular and extensible way. In the following, we briefly describe the architecture shown in figure 3, from bottom to top.

4.2.1 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)*[11] and the *Transfer Object (TO)*[11] 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 storing managers, there is the *log storing manager* 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 storing manager allows us to fine trace the provenance of each piece of data from its entrance in the system to every further processing on it.

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

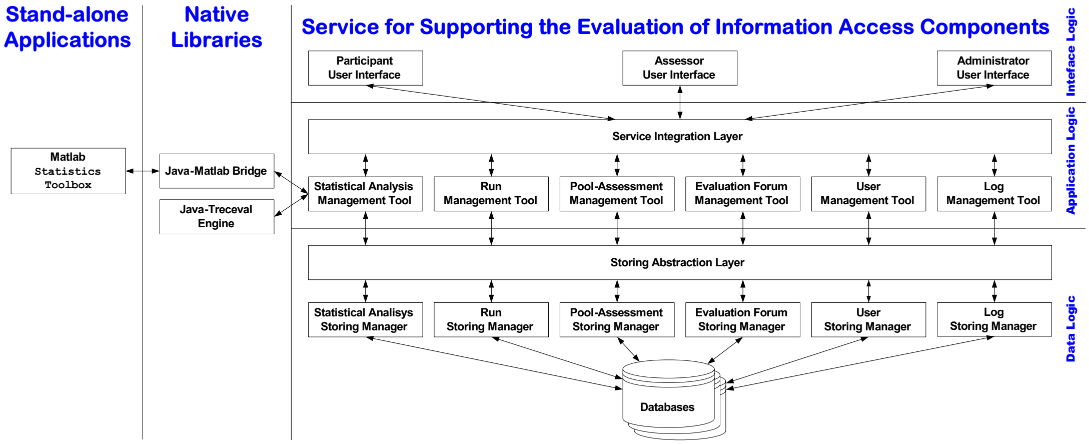

Stand-alone Applications
Native Libraries
Service for Supporting the Evaluation of Information Access Components
Inteface Logic
Application Logic
Data Logic
Participant User Interface
Assessor User Interface
Administrator User Interface
Service Integration Layer
Matlab Statistics Toolbox
Java-Matlab Bridge
Java-Treceval Engine
Statistical Analysis Management Tool
Run Management Tool
Pool-Assessment Management Tool
Evaluation Forum Management Tool
User Management Tool
Log Management Tool
Storing Abstraction Layer
Statistical Analisys Storing Manager
Run Storing Manager
Pool-Assessment Storing Manager
Evaluation Forum Storing Manager
User Storing Manager
Log Storing Manager
Databases

Figure 3. Service architecture for supporting evaluation of information access components.

Finally, on top of the various "storing managers" there is the *Storing Abstraction Layer (SAL)* which hides the details about the storage management to the upper layers. In this way, the addition of a new "storing manager" is totally transparent for the upper layers.

4.2.2 Application Logic

The application logic layer deals with the flow of operations within *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)*. It provides a set of tools capable of managing high-level tasks, such as experiment submission, pool assessment, statistical analysis of an experiment.

For example, the *Statistical Analysis Management Tool (SAMT)* offers the functionalities needed to conduct a statistical analysis on a set of experiments. In order to ensure comparability and reliability, the SAMT makes uses of well-known and widely used tools to implement the statistical tests, so that everyone can replicate the same test, even if he has no access to the service. In the architecture, the MATLAB Statistics Toolbox[12] 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 softwares could require different solutions. As an additional example aimed at wide comparability and acceptance of the tools, a further library provides an interface for our service towards the `trec_eval` package[13]. `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 *Service Integration Layer (SIL)* provides the interface logic layer with a uniform and integrated access to the various tools. As we noticed in the case of the SAL, thanks to the SIL also the addition of new tools is transparent for the interface logic layer.

4.2.3 Interface Logic

It is the highest level of the architecture, and it is the access point for the user to interact with the system. It provides specialised *User Interfaces (UIs)* for different types of users, that are the participants, the assessors, and the administrators. Note that, thanks to the abstraction provided by the application logic layer, different kind of UIs can be provided, either stand-alone applications or Web-based applications.

5 Running System

The proposed software infrastructure has been implemented in a prototype, called *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* [13, 16], and it has been tested in the context of the CLEF 2005 and 2006 evaluation campaigns. The prototype provides support for:

- the management of an evaluation forum: the track set-up, the harvesting of documents, the management of the subscription of participants to tracks;
- the management of submission of experiments, the collection of metadata about experiments, and their validation;
- the creation of document pools and the management of relevance assessment;
- common statistical analysis tools for both organizers and participants in order to allow the comparison of the experiments;

[12] `http://www.mathworks.com/products/statistics/`

[13] `ftp://ftp.cs.cornell.edu/pub/smart/`

- common tools for summarizing, producing reports and graphs on the measured performances and conducted analyses;
- common *eXtensible Markup Language (XML)* format for exchanging data between organizers and participants.

DIRECT was successfully adopted during the CLEF 2005 campaign. It was used by nearly 30 participants spread over 15 different nations, who submitted more than 530 experiments; then 15 assessors assessed more than 160,000 documents in seven different languages, including Russian and Bulgarian which do not have a latin alphabet. During the CLEF 2006 campaign, it has been used by nearly 75 participants spread over 25 different nations, who have submitted around 570 experiments; 40 assessors assessed more than 198,500 documents in nine different languages. DIRECT was then used for producing reports and overview graphs about the submitted experiments [14].

DIRECT has been developed by using the Java[14] programming language, which ensures great portability of the system across different platforms. We used the PostgreSQL[15] *DataBase Management System (DBMS)* for performing the actual storage of the data. Finally, all kinds of UI in DIRECT are Web-based interfaces, which make the service easily accessible to end-users without the need of installing any kind of software. These interfaces have been developed by using the Apache STRUTS[16] framework, an open-source framework for developing Web applications.

Figure 4 shows the user interface for the management of the submitted experiments by the participant.

Figure 5 shows the user interface offered to the assessor for making the relevance assessments.

Finally, figure 6 shows some of the performance measurements and descriptive statistics available to the participants.

6 Conclusions

This study has addressed the methodology currently adopted for the experimental evaluation in the IR field, and it has proposed to extend it including a proper management, curation, archiving, and enrichment of the scientific data that are produced while conducting an experimental evaluation in the context of an evaluation campaign.

We have discussed several possible ways of extending the current methodology, and to positively contribute to an organic and systematic extension of it, we have presented a conceptual model and an effective architecture for developing an innovative software infrastructure to support the course of an evaluation campaign. The prototype DIRECT, that implements both the conceptual model and the architecture, has been introduced. DIRECT has shown to be robust in its use during CLEF 2005 and 2006 evaluation campaigns. On the basis of the experience gained, we are enhancing the proposed conceptual model and architecture, in order to further enrich both the model and the prototype to widen the support to researchers that are going to participate in the next CLEF 2007 campaign.

Acknowledgements

The work reported in this paper has been partially supported by the DELOS Network of Excellence on Digital Libraries, as part of the Information Society Technologies (IST) Program of the European Commission (Contract G038-507618). The work was also partially supported by the SAPIR project, as part of the Information Society Technologies (IST) Program of the European Commission (Contract IST-045128).

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

We thank the anonymous referees for their fruitful comments and careful reading that improved this article.

References

[1] S. Abiteboul, R. Agrawal, P. Bernstein, M. Carey, S. Ceri, B. Croft, D. DeWitt, M. Franklin, H. Garcia-Molina, D. Gawlick, J. Gray, L. Haas, A. Halevy, J. Hellerstein, Y. Ioannidis, M. Kersten, M. Pazzani, M. Lesk, D. Maier, J. Naughton, H.-J. Schek, T. Sellis, A. Silberschatz, M. Stonebraker, R. Snodgrass, J. D. Ullman, G. Weikum, J. Widom, and S. Zdonik. The Lowell Database Research Self-Assessment. *Communications of the ACM (CACM)*, 48(5):111–118, 2005.

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

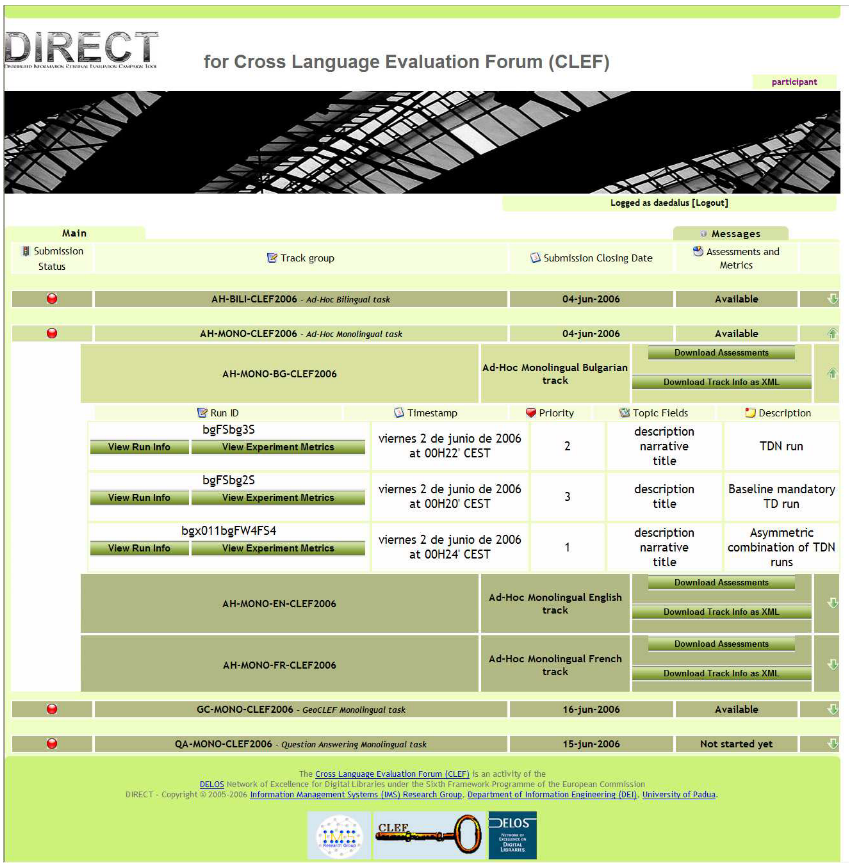
[3] M. Agosti, G. M. Di Nunzio, and N. Ferro. A Data Curation Approach to Support In-depth Evaluation Studies. In F. C. Gey, N. Kando, C. Peters, and C.-Y. Lin, editors, *Proc. International Workshop on New Directions in Multilingual Information Access (MLIA 2006)*, pages 65–68. `http://ucdata.berkeley.edu/sigir2006-mlia.htm` [last visited 2006, October 2], 2006.

[4] M. Agosti, G. M. Di Nunzio, and N. Ferro. Scientific Data of an Evaluation Campaign: Do We Properly Deal With Them? In A. Nardi, C. Peters, and J. L. Vicedo, editors, *Working Notes for the CLEF 2006 Workshop*. `http://www.clef-campaign.org/2006/working_notes/workingnotes2006/agostiCLEF2006.pdf` [last visited 2006, October 2], 2006.

[14] `http://java.sun.com/`

[15] `http://www.postgresql.org/`

[16] `http://struts.apache.org/`


DIRECT
for Cross Language Evaluation Forum (CLEF)
participant
Logged as daedalus [Logout]
Main
Messages

Submission Status	Track group	Submission Closing Date	Assessments and Metrics
●	AH-BILI-CLEF2006 - *Ad-Hoc Bilingual task*	04-jun-2006	Available
●	AH-MONO-CLEF2006 - *Ad-Hoc Monolingual task*	04-jun-2006	Available

AH-MONO-BG-CLEF2006	Ad-Hoc Monolingual Bulgarian track	Download Assessments Download Track Info as XML

Run ID	Timestamp	Priority	Topic Fields	Description
bgFSbg3S View Run Info · View Experiment Metrics	viernes 2 de junio de 2006 at 00H22' CEST	2	description narrative title	TDN run
bgFSbg2S View Run Info · View Experiment Metrics	viernes 2 de junio de 2006 at 00H20' CEST	3	description title	Baseline mandatory TD run
bgx011bgFW4FS4 View Run Info · View Experiment Metrics	viernes 2 de junio de 2006 at 00H24' CEST	1	description narrative title	Asymmetric combination of TDN runs

AH-MONO-EN-CLEF2006	Ad-Hoc Monolingual English track	Download Assessments Download Track Info as XML
AH-MONO-FR-CLEF2006	Ad-Hoc Monolingual French track	Download Assessments Download Track Info as XML

●	GC-MONO-CLEF2006 - *GeoCLEF Monolingual task*	16-jun-2006	Available
●	QA-MONO-CLEF2006 - *Question Answering Monolingual task*	15-jun-2006	Not started yet

The Cross Language Evaluation Forum (CLEF) is an activity of the
DELOS Network of Excellence for Digital Libraries under the Sixth Framework Programme of the European Commission
DIRECT - Copyright © 2005-2006 Information Management Systems (IMS) Research Group, Department of Information Engineering (DEI), University of Padua.

Figure 4. Participant user interface for the management of the experiments.

[5] W. L. Anderson. 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, December 2004.

[6] T. Berners-Lee. Universal Resource Identifiers in WWW. RFC 1630, June 1994.

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

[8] G. Booch, J. Rumbaugh, and I. Jacobson. *The Unified Modeling Language User Guide*. Addison-Wesley, Reading (MA), USA, 1999.

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

[10] C. Buckley and D. Harman. Reliable Information Access Final Workshop Report. `http://nrrc.mitre.org/NRRC/Docs_Data/RIA_2003/ria_final.pdf` [last visited 2007, January 4], January 2004.

[11] C. W. Cleverdon. The Cranfield Tests on Index Languages Devices. In K. Spack Jones and P. Willett, editors, *Readings in Information Retrieval*, pages 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, California, USA, 1997.

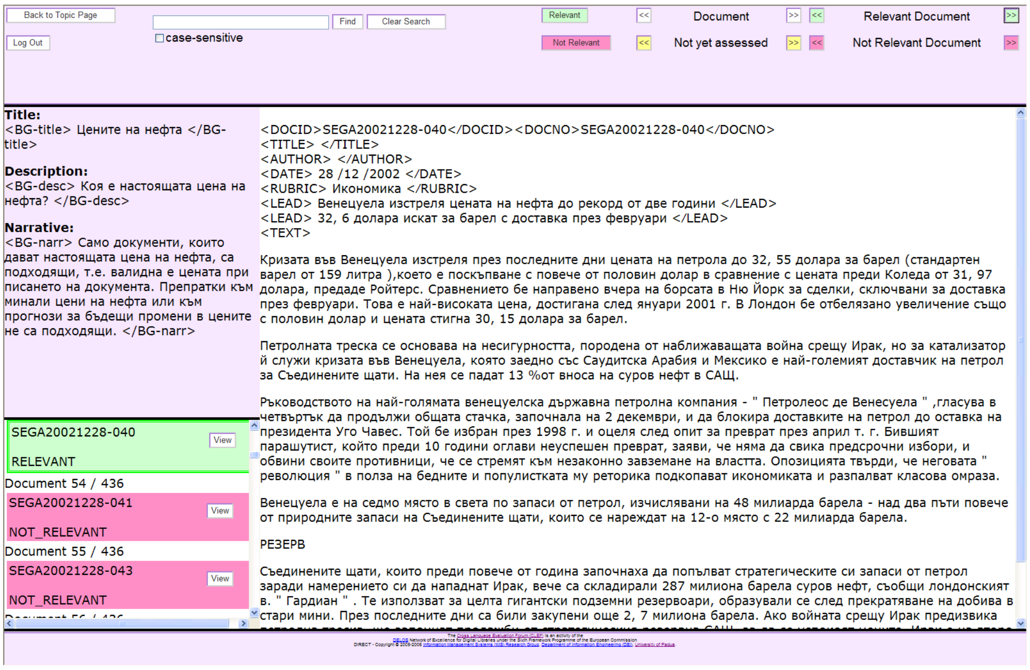
[12] W. B. Croft. Combining Approaches to Information Retrieval. In W. B. Croft, editor, *Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval*, pages 1–36. Kluwer Academic Publishers, Norwell (MA), USA, 2000.

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

[14] G. M. Di Nunzio and N. Ferro. Appendix A: Results of the Ad-hoc Bilingual and Monolingual Tasks. In A. Nardi, C. Peters, and J. L. Vicedo, editors, *Working Notes for the CLEF 2006 Workshop*. `http://www.clef-campaign.org/2006/working_notes/workingnotes2006/Appendix_Ad-Hoc.pdf` [last visited 2006, October 2], 2006.

[15] G. M. Di Nunzio and N. Ferro. Queries and Relevance Assessments: The Right Context for the Right Topic. In *Proc. First International Workshop on Adaptive Information Retrieval (AIR)*. `http://www.dcs.gla.ac.uk/workshops/air/` [last visited 2007, January 4], October 2006.

[16] G. M. Di Nunzio and N. Ferro. Scientific Evaluation of a DLMS: a service for evaluating informa-



Title:
<BG-title> Цените на нефта </BG-title>

Description:
<BG-desc> Коя е настоящата цена на нефта? </BG-desc>

Narrative:
<BG-narr> Само документи, които дават настоящата цена на нефта, са подходящи, т.е. валидна е цената при писането на документа. Препратки към минали цени на нефта или към прогнози за бъдещи промени в цените не са подходящи. </BG-narr>

SEGA20021228-040
RELEVANT
Document 54 / 436
SEGA20021228-041
NOT_RELEVANT
Document 55 / 436
SEGA20021228-043
NOT_RELEVANT

```
<DOCID>SEGA20021228-040</DOCID><DOCNO>SEGA20021228-040</DOCNO>
<TITLE> </TITLE>
<AUTHOR> </AUTHOR>
<DATE> 28 /12 /2002 </DATE>
<RUBRIC> Икономика </RUBRIC>
<LEAD> Венецуела изстреля цената на нефта до рекорд от две години </LEAD>
<LEAD> 32, 6 долара искат за барел с доставка през февруари </LEAD>
<TEXT>
```

Кризата във Венецуела изстреля през последните дни цената на петрола до 32, 55 долара за барел (стандартен варел от 159 литра),което е поскъпване с повече от половин долар в сравнение с цената преди Коледа от 31, 97 долара, предаде Ройтерс. Сравнението бе направено вчера на борсата в Ню Йорк за сделки, сключвани за доставка през февруари. Това е най-високата цена, достигана след януари 2001 г. В Лондон бе отбелязано увеличение също с половин долар и цената стигна 30, 15 долара за барел.

Петролната треска се основава на несигурността, породена от наближаващата война срещу Ирак, но за катализатор й служи кризата във Венецуела, която заедно със Саудитска Арабия и Мексико е най-големият доставчик на петрол за Съединените щати. На нея се падат 13 %от вноса на суров нефт в САЩ.

Ръководството на най-голямата венецуелска държавна петролна компания - " Петролеос де Венесуела " ,гласува в четвъртък да продължи общата стачка, започнала на 2 декември, и да блокира доставките на петрол до оставка на президента Уго Чавес. Той бе избран през 1998 г. и оцеля след опит за преврат през април т. г. Бившият парашутист, който преди 10 години оглави неуспешен преврат, заяви, че няма да свика предсрочни избори, и обвини своите противници, че се стремят към незаконно завземане на властта. Опозицията твърди, че неговата " революция " в полза на бедните и популистката му реторика подкопават икономиката и разпалват класова омраза.

Венецуела е на седмо място в света по запаси от петрол, изчислявани на 48 милиарда барела - над два пъти повече от природните запаси на Съединените щати, които се нареждат на 12-о място с 22 милиарда барела.

РЕЗЕРВ

Съединените щати, които преди повече от година започнаха да попълват стратегическите си запаси от петрол заради намерението си да нападнат Ирак, вече са складирали 287 милиона барела суров нефт, съобщи лондонският в. " Гардиан " . Те използват за целта гигантски подземни резервоари, образували се след прекратяване на добива в стари мини. През последните дни са били закупени още 2, 7 милиона барела. Ако войната срещу Ирак предизвика

The Cross Language Evaluation Forum (CLEF) is an activity of the
DELOS Network of Excellence for Digital Libraries under the Sixth Framework Programme of the European Commission
DIRECT - Copyright © 2005-2006 Information Management Systems (IMS) Research Group, Department of Information Engineering (DEI), University of Padua.

Figure 5. Assessor user interface for performing the relevance assessments.

tion access components. In J. Gonzalo, C. Thanos, M. F. Verdejo, and R. C. Carrasco, editors, *Proc. 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006)*, pages 536–539. Lecture Notes in Computer Science (LNCS) 4172, Springer, Heidelberg, Germany, 2006.

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

[18] European Commission Information Society and Media. i2010: Digital Libraries. `http://europa.eu.int/information_society/activities/digital_libraries/doc/brochures/dl_brochure_2006.pdf` [last visited 2006, October 2], October 2006.

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

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

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

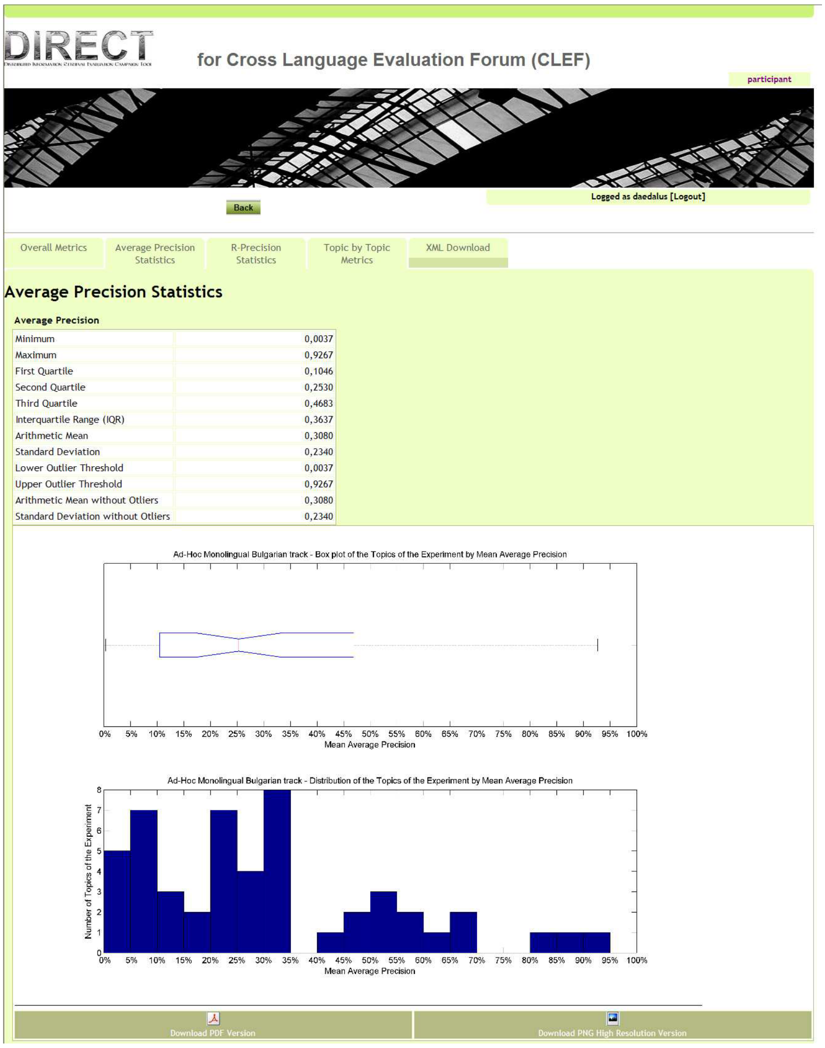
[22] Y. Ioannidis, D. Maier, S. Abiteboul, P. Buneman, S. Davidson, E. A. Fox, A. Halevy, C. Knoblock, F. Rabitti, H.-J. Schek, and G. Weikum. Digital library information-technology infrastructures. *International Journal on Digital Libraries*, 5(4):266–274, 2005.

[23] K. Kishida, K.-h. Chen, S. Lee, D. Kuriyama, N. Kando, H.-H. Chen, and S. H. Myaeng. Overview of CLIR Task at the Fifth NTCIR Workshop. In N. Kando and M. Takaku, editors, *Proc. of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*. `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/CLIR/NTCIR5-OV-CLIR-KishidaK.pdf` [last visited 2007, January 4], 2005.

[24] P. Lord and A. Macdonald. *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` [last visited 2006, October 2], 2003.

[25] National Science Board. *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/` [last visited 2007, January 3], September 2005.

[26] NISO. *ANSI/NISO Z39.88 - 2004 – The OpenURL Framework for Context-Sensitive Services*. National Information Standards Organization (NISO).


DIRECT
for Cross Language Evaluation Forum (CLEF)
participant
Logged as daedalus [Logout]
Back
Overall Metrics
Average Precision Statistics
R-Precision Statistics
Topic by Topic Metrics
XML Download
Ad-Hoc Monolingual Bulgarian track - Box plot of the Topics of the Experiment by Mean Average Precision
0% 5% 10% 15% 20% 25% 30% 35% 40% 45% 50% 55% 60% 65% 70% 75% 80% 85% 90% 95% 100%
Mean Average Precision
Ad-Hoc Monolingual Bulgarian track - Distribution of the Topics of the Experiment by Mean Average Precision
Number of Topics of the Experiment
0 1 2 3 4 5 6 7 8
0% 5% 10% 15% 20% 25% 30% 35% 40% 45% 50% 55% 60% 65% 70% 75% 80% 85% 90% 95% 100%
Mean Average Precision
Download PDF Version
Download PNG High Resolution Version

Average Precision Statistics

Average Precision

Minimum	0,0037
Maximum	0,9267
First Quartile	0,1046
Second Quartile	0,2530
Third Quartile	0,4683
Interquartile Range (IQR)	0,3637
Arithmetic Mean	0,3080
Standard Deviation	0,2340
Lower Outlier Threshold	0,0037
Upper Outlier Threshold	0,9267
Arithmetic Mean without Otliers	0,3080
Standard Deviation without Otliers	0,2340

Figure 6. Performance measurements and descriptive statistics available to participants.

http://www.niso.org/standards/standard_detail.cfm?std_id=783 [last visited 2006, October 2], April 2005.

[27] OAI. The Open Archives Initiative Protocol for Metadata Harvesting – Version 2.0. http://www.openarchives.org/OAI/openarchivesprotocol.html [last visited 2006, October 2], October 2004.

[28] N. Paskin. Digital Object Identifiers for Scientific Data. *Data Science Journal*, 4:12–20, April 2005.

[29] N. Paskin, editor. *The DOI Handbook – Edition 4.4.0*. International DOI Foundation (IDF). http://dx.doi.org/10.1000/186 [last visited 2006, October 2], September 2006.

[30] J. Rumbaugh, I. Jacobson, and G. Booch. *The Unified Modeling Language Reference Manual*. Addison-Wesley, Reading (MA), USA, 1999.

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

[32] E. M. Voorhees. Overview of the TREC 2005 Robust Retrieval Track. In E. M. Voorhees and L. P. Buckland, editors, *The Fourteenth Text REtrieval Conference Proceedings (TREC 2005)*. http://trec.nist.gov/pubs/trec14/t14_proceedings.html [last visited 2006, October 2], 2005.

[33] Working Group on Data for Science. *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 [last visited 2007, January 3], December 2006.

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

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