Towards an infrastructure for digital library performance evaluation

Maristella Agosti and Nicola Ferro

Introduction

A few years ago, Ioannidis et al. (2005) observed that 'Digital Library (DL) development must move from an art to a science' to give rise to digital library systems (DLS) based on reliable and extensible services and of proven quality.

Given the growth of the DLS and the need for a scientific approach in developing them, proper evaluation methodologies are needed to assess their performance along different dimensions. Such evaluation methodologies should not be perceived as something external to the design and development process of these complex systems, but rather they should be tightly integrated into it.

Moreover, the actual evaluation of a DLS is a scientific activity where the outcomes, such as performance analyses and measurements, constitute a kind of scientific data that must be properly taken into consideration and used for the design and development of DLS components and services.

Interestingly enough, this line of reasoning highlights a kind of intrinsic circularity when digital library development is viewed as a science. Indeed, achieving the necessary levels of reliability and effectiveness for such large-scale DLS calls for extensive use of evaluation methodologies which, as a result, produce a considerable amount of scientific data. These data should, in turn, be managed by a DLS that supports their enrichment, interpretation and preservation in order to yield the expected positive feedback on DLS design and development. Indeed, Ioannidis et al. (2005) argue that a DLS should support information enrichment

and that 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'. In addition, they observe that citation, intended as the possibility of explicitly mentioning and making reference to portions of a given digital object, should also be part of the information enrichment strategies supported by a DLS.

Therefore, the scientific development of the contents of a digital library would greatly benefit from the exploitation of a DLS for scientific data which provides the tools for enriching, citing, interpreting, and preserving the scientific data produced during the design and development process.

This type of DLS shall be known as a *scientific reflection DLS*, as it deals with scientific data, information, and interpretations about the design and development of another DLS. Indeed, the term *reflection* means both 'the act of reflecting or the state of being reflected' and 'careful or long consideration or thought' (Hanks, 1979). Here, the first meaning of reflection illustrates the capability of a scientific reflection DLS to show the evaluation outcomes of a target DLS, while the second meaning of reflection implies that a scientific reflection DLS provides support for designing and developing a target DLS.

The evaluation of a DLS is a non-trivial issue which should analyse different aspects, such as architecture, information access and extraction capabilities, management of multimedia content, interaction with users, and so on (Führ et al., 2001, 2007). As there are so many aspects to take into consideration, the scientific reflection DLS should be constituted by different and cooperating services, each focused on supporting the evaluation of one of the abovementioned aspects.

Attention will therefore be focused on the current evaluation methodologies for assessing the performances of the information access and extraction components of a DLS, which deal with the indexing, search and retrieval of documents in response to a user's query. In particular, these methodologies will be investigated to see whether they meet the requirements of information enrichment necessary for the scientific development of DLS.

This chapter focuses on the revision of current evaluation methodologies in order to adapt to the new way of thinking about DLS development. Furthermore, the outcomes of this revision process are applied to the design and development of a service capable of supporting the evaluation of the information access and extraction components of a DLS. This service is intended to be part of a wider scientific reflection DLS, which covers different aspects of DLS evaluation, and represents a

first step towards the creation of a comprehensive evaluation infrastructure for assessing DLS performances from different viewpoints. Finally, the chapter describes a running prototype of this DLS service called DIRECT (*direct.dei.unipd.it*) which implements the proposed revision of current evaluation methodologies.

Conceptual framework for the evaluation of DLS information access components

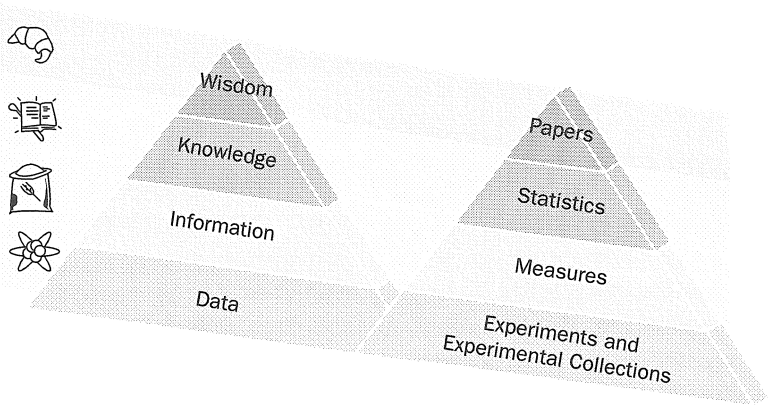
The current approach for laboratory evaluation of information access systems relies on the Cranfield methodology, which makes use of experimental collections (Cleverdon, 1997). An experimental collection is a triple $C = (D, T, J)$, where: D is a set of documents, also called collection of documents; T is a set of topics, which expresses the user's information needs and from which the actual queries are derived; and J is a set of relevance judgments, i.e. for each topic $t \in T$ and for each document $d \in D$ it is determined whether or not d is relevant to t.

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 is a scientific activity and, as such, its outcomes are different kinds of valuable scientific data. Therefore, the experiments themselves represent the primary scientific data and the starting point of the investigation. Using the experimental data, different performance measurements are produced, such as precision and recall, which are standard measures to evaluate the performances of an information access component for a given experiment. Starting from these performance measurements, descriptive statistics can be computed, such as mean or median, to summarise the overall performances achieved by an experiment or by a collection of experiments. Finally, hypothesis tests and other statistical analyses can be performed to conduct an in-depth analysis and comparison over a set of experiments.

The abovementioned scientific data can be framed in the context of the data, information, knowledge, wisdom (DIKW) hierarchy (Ackoff, 1989; Zeleny, 1987), represented in Figure 6.1:

Figure 6.1 The DIKW hierarchy with respect to the experimental evaluation



- At the *data layer* there are raw, basic elements, partial and atomised, which have little meaning by themselves and no significance beyond their immediate 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 experiments and the experimental collections correspond to the 'data level' in the hierarchy, as they represent the raw, basic elements needed for any further investigation and have little meaning by themselves. Indeed, without a relationship with the experimental collection to which the experiment pertains, an experiment and the associated results are of limited value, as these data constitute the basis for any subsequent computation.
- 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 performance measurements correspond to the 'information level' in the hierarchy, as they are the result of computations and processing on the data, so that a meaning is associated with the data by way of some kind of relational connection. For example, precision and recall measures are obtained by relating the results of an experiment with the relevance judgments J.

- 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 descriptive statistics and the hypothesis tests correspond to the 'knowledge level' in the hierarchy, as they are a further elaboration of the information carried by the performance measurements and provide some insights about the experiments.
- The *wisdom level* provides interpretation, explanation and formalisation 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. 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, as they provide interpretation, explanation, and formalisation of the content of the previous levels.

As observed by Zeleny:

> 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 judgment. (Zeleny, 1987)

This observation also fits the case of the experimental evaluation. Indeed, experiments (data) and performance measurements (information) are usually generated in an automatic way by information access components, programs and tools for assessing performances. On the other hand, statistical analyses (knowledge) and models and algorithms (wisdom) require the deep involvement of researchers in order to be conducted and developed.

This view of the 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 (Agosti et al., 2007a; Dussin and Ferro, 2008a).

This question is made more compelling by the fact that when dealing with scientific data, 'the lineage (provenance) of the data must be tracked, as 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' (Abiteboul et al., 2005). Moreover, Ioannidis et al. (2005) point 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 (Agosti et al., 2007b). Finally, the National Science Board (2005) 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 the degree to which 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.

User requirements analysis

Different types of actors are involved in an evaluation campaign (Dussin and Ferro, 2007):

- 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 performance of his experiments and overall indicators that allow his experiments and results to be compared with those submitted by other participants. Moreover, he should have the possibility of properly citing his experiments and other information resources and having 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 that provides insights about the conducted experiments. Moreover, he should have the possibility of properly citing the accessed information resources and having a citation correctly resolved to the corresponding information resources.
- The *organiser* 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 characterise 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.

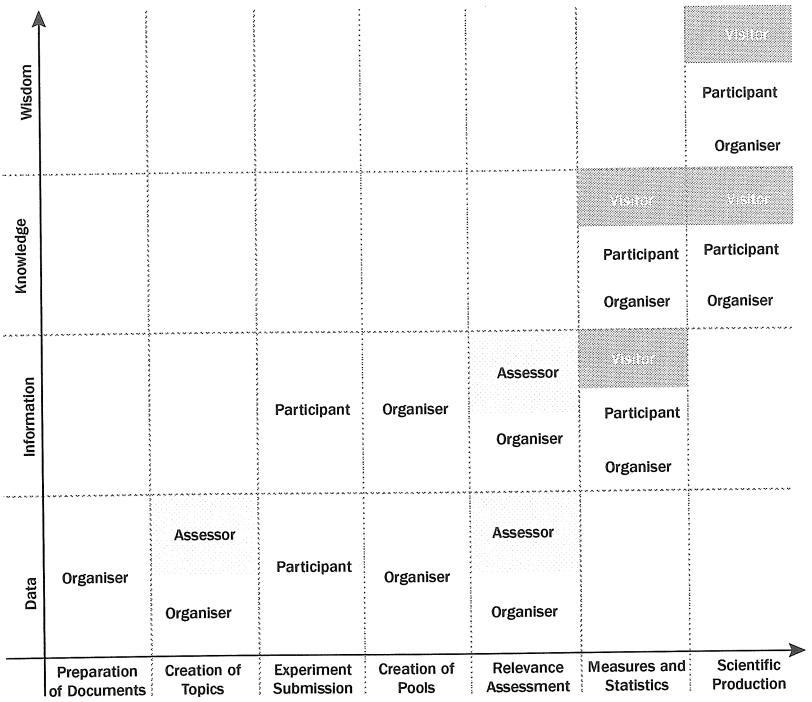
These actors interact together in various ways during the course of an evaluation campaign and contribute differently to the DIKW hierarchy discussed above.

Figure 6.2 shows 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 the campaign comes into full swing, there is a progressive movement from data to wisdom and also the number of actors involved and their interaction grows.

The elements shown in Figure 6.2 will now be examined in more detail:

- *Acquisition and preparation of documents*: the organisers 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. Organisers need an interface that allows them to upload the collections of documents, which can be in diverse media, into the DLS to make them available to participants and assessors.
- *Creation of topics*: the organisers and the assessors cooperate to create the topics for the test collection. (*Topic* is the term we adopt for

Figure 6.2 Relationship between the DIKW hierarchy, the different types of actors, and the main steps of an evaluation campaign



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 particular tasks for which they will be used.) For each topic, this step usually requires preparing a first draft of the topic and searching the set of documents to verify that there are relevant documents for that topic. The topic is then 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. Organisers 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 the contents of the topic. Note that topics are created by inspecting the documents, which in a sense are a kind of data more

basic than the topics. This fact is also reflected in the user interface, which needs to support more complex tasks that 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 of relevance for each topic and represents the output of the execution of the information retrieval system (IRS) developed by the participant. The experiments are part of the data 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, and in a sense represent a kind of more complex data with respect to them. This fact is also reflected in the user interface, which provides support for checking the correctness of the experiments with respect to topics and documents.
- *Creation of pools*: the organisers 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 to determine their actual relevance. The pools are midway between data and information, as they are still quite raw elements but represent a first form of processing of the experiments. Organisers need an interface that allows them to select and sample the documents to be inserted in the pool and to see dynamically how the pools change when the selection criteria are modified to determine the best strategy for creating the pools. This hybrid nature of the pools between data and information is also reflected 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 organisers and the assessors cooperate to assess each document in the pool with respect to the topic, i.e. for determining whether or not the document is relevant for the given topic. As in the case of the pools, the relevance judgments are midway between data and information, as they are raw elements which constitute an experimental collection but represent human-added information about the relationship between the topics and documents

of an experiment. Organisers need an interface that allows them to set up and monitor the relevance assessment process and to publish the relevance judgments 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 for discussion in the event of topics that may be difficult or ambiguous to assess. This dual nature of the relevance assessment between data and information is also reflected in the user interface, which explicitly requests the assessors to enter a human judgment (relevant or not relevant) about the relationship between a document and a topic.

- *Measures and statistics*: the organisers exploit the relevance assessments to compute the performance measures and plots about each experiment submitted by a participant. These measurements are then 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, as they are the results of data processing; descriptive statistics and hypothesis tests are knowledge, as they provide further insights into the meaning of the obtained performances. Organisers need an interface that allows them to perform all the computations and statistical analyses that are needed. Participants and visitors need an interface that gives access and presents performance measurements, plots, descriptive statistics and statistical analyses in a meaningful way in order to facilitate their comprehension and interpretation.
- *Scientific production*: both organisers 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 even after the conclusion of the campaign, as 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 organisers and the participants but also external visitors may exploit the information resources produced during the evaluation campaign to carry out their research activity. As explained above, the outcome of this process is wisdom. Organisers, 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 the needs of the users involved in a large-scale evaluation campaign are 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 internationalisation and localisation of the interface in order to lower language and cultural barriers.

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 in the systematic definition of a proper methodology for managing and promoting the use of data.

The e-Science Data Curation Report gives the following definition of data curation:

> 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. (Lord and MacDonald, 2003)

This definition implies the need to consider 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. Such reasons include, for example, reuse of data for new research, including collection-based research to generate new science; retention of unique observational data which cannot be recreated; retention of expensively generated data which is cheaper to maintain than to regenerate; enhancing existing data available for research projects; and validating published research results.

As a concrete example in the field of information retrieval, consider the data fusion problem (Croft, 2000), 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 use other researchers' results to develop their merging algorithms. In 2005, for example, the CLEF, the European initiative for the evaluation of multilingual information access components (*www.clef-campaign.org*), ran a multilingual merging track to provide results from experiments it had run in 2003 for participants to use as data for their merging algorithms (Di Nunzio et al., 2006). It is clear that such researchers would benefit from a data curation strategy which could promote the reuse of existing data and allow data fusion experiments to be traced back to the original 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 modelling, managing and curating the scientific data produced during an evaluation campaign. The following sections present a discussion of some key points that were taken into consideration when designing the DIRECT DLS and which extend the current evaluation methodology.

Conceptual model

The definition of experimental collection does not take into consideration any kind of conceptual model (Tsichritzis and Lochovsky, 1982), neither of 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 for making 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 an appropriate logical model in order to manage the information of an evaluation campaign by using robust data management technology. From the conceptual model, appropriate data formats can also be derived for exchanging information among organisers and participants.

The conceptual model is built around five main modelling areas:

- *evaluation campaign*: this 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 organised, the subscription of the participants to the tracks, and the topics of each track;
- *collection*: this concerns the different collections made available by an evaluation forum; each collection can be organised 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*: this regards the experiments submitted by the participants and the evaluation metrics computed on those experiments, such as precision and recall;
- *pool/relevance assessment*: this 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 to which the experiments belong;
- *statistical analysis*: this 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 entity in the conceptual model has the possibility of being enriched with various metadata objects 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 in turn be annotated with other metadata objects, so that is possible to have a chain of nested metadata describing a given object.

Metadata

Anderson (2004) 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'. As there is no conceptual model for an experimental collection, appropriate metadata schemes are also lacking. Consider that there are almost no metadata:

- to describe a collection of documents D – useful metadata would concern, at least, the creator, the creation date, a description, the context of the collection, and how the collection has been created;
- about the topics T – useful metadata would describe the creators and the creation date, how the creation process has taken place, if there were any issues, what documents the creators found relevant for a given topic, and so on;
- to describe the relevance judgments J – examples of such metadata concern creators and the creation date, the criteria leading to the creation of the relevance judgments, the problems faced by the assessors when dealing with difficult topics.

The situation is a little 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, and whether the query has been automatically or manually constructed from the topics. The Text Retrieval Conference (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, and what tunings have been carried out.

A good attempt in this direction is the Reliable Information Access Workshop (Harman and Buckley, 2004), organised by the US National Institute of Standards and Technology 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 also hold 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 judgments were used to compute a performance measure, and when the performance measure was computed. Similar metadata could also be useful for descriptive statistics and statistical analyses.

All this additional information can provide useful hints about the system models as well as the context of the evaluation. The context is not simply the track or specific experiments, as more information might be needed, 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 higher-level research activities, such as assessing the reliability of information-retrieval experiments (Zobel, 1998).

Unique identification mechanism

The lack of a conceptual model causes another relevant consequence: there is no common mechanism for uniquely identifying 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 judgments, experiments and statistical analyses.

The absence of a mechanism to uniquely identify and reference the digital objects of an evaluation campaign prevents the direct citation of that digital object. Indeed, as recognised by Lord and MacDonald (2003), the possibility of citing scientific data and elaborating on it further is an effective way of making scientists and researchers an active part of the digital curation process. Moreover, this opportunity would strengthen the passing from data to wisdom because experimental collections and experiments would become as citable and accessible as any other item in the reference list of a paper.

Among the various identification solutions that have been proposed, the 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 uniform resource identifier (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 and provides facilities for resolving the identifiers (Paskin, 2006). This means that it enables direct access to each identified

digital object starting from its identifier, in this way giving an interested researcher direct access to the referenced digital object together with all the information concerning it. Finally, the DOI constitutes a valuable possibility for identifying and referencing digital objects of an evaluation campaign, as there have already been successful attempts to apply it to scientific data, and it also makes possible the association of metadata with identified digital objects, as observed by Brase (2004) and Paskin (2005).

The DOI has therefore been adopted as a unique identification mechanism and DOIs are registered for different information resources of an evaluation campaign in accordance with mEDRA (*www.medra.org*).

Statistical analyses

Hull (1993) points out that, in order to evaluate retrieval performances, not only does one need 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. Furthermore, participants may conduct statistical analyses on their own experiments by using either ad hoc packages, such as IR-STAT-PAK, or generally available software tools with statistical analysis capabilities, like R, SPSS or MATLAB. However, the choice of whether or not to perform a statistical analysis 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 analyse the data, or different choices and approximations for the various parameters of the same statistical test can be made.

Therefore, support and guidance to participants have been provided to adopt a more uniform way of performing statistical analyses on their own experiments. Indeed, not only can participants benefit from standard experimental collections which make their experiments comparable, 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 stated above, 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 provides the means for further elaborating and interpreting the experimental results. Nevertheless, the current methodology does not require any particular coordination or synchronisation between the basic scientific data and the analyses on them, which are treated as almost separate items. In contrast, researchers could greatly benefit from an integrated vision of them, 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 of them.

Statistical analyses concerning the performances of each experiment are carried out, 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. Thus, through the provision of tools to ease their work, participants are encouraged to conduct in-depth analyses of their results. Moreover, statistical analyses and hypothesis tests are conducted, 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 trend and behaviour for a given task.

Architecture of the DIRECT system

As a 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, organisers and visitors, who need to have access to different kinds of features and capabilities;

- to support internationalisation and localisation – the application needs to be able to adapt to the language of the user and their country or culturally dependent data, such as dates and currencies.

Figure 6.3 shows the architecture of the proposed service. It consists of three layers – data, application and interface logic layers – in order to achieve improved modularity and to 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 manner (Agosti et al., 2007a; Dussin and Ferro, 2008b).

In the following, a brief description of the architecture shown in Figure 6.3 is given from bottom to top.

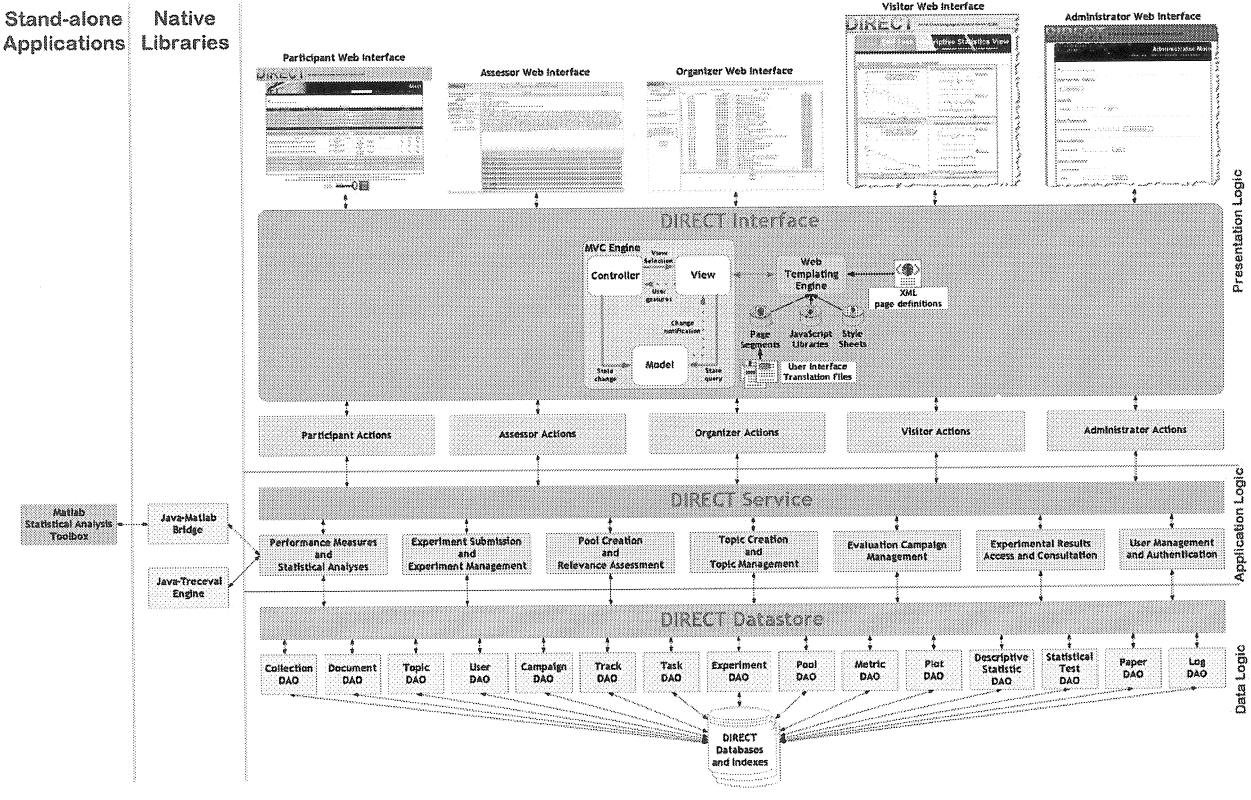
Data logic

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. The data access object and the transfer object design patterns have been adopted (*java.sun.com/blueprints/corej2eepatterns/Patterns/*). The data access object 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 data access object to adapt to different storage schemes without affecting the upper layers.

In addition to the other data access objects, there is the log data access object which accurately traces both system and user events. It captures information such as the user name, the IP address of the connecting host, the action that has been invoked by the user, the messages exchanged between the components of the system in order to carry out the requested action, any error condition, and so on. Thus, besides providing a log of the system and user activities, the log data access object enables the provenance of each piece of data to be accurately traced from its entrance in the system to every further processing of it.

Finally, on top of the various data access objects 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 data access object is totally transparent for the upper layers.

Figure 6.3 Architecture of the DIRECT system



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 use of well-known and widely-used tools to implement the statistical tests, so that everyone can replicate the same test, even if they have no access to the service. In the architecture, the MATLAB Statistics Toolbox has been adopted, as MATLAB is a leader application in the field of numerical analysis which employs state-of-the-art algorithms, although other software could have been used as well. In the case of MATLAB, an additional library is needed to allow the service to access MATLAB in a programmed way; other applications might require different solutions. A further library provides an interface for the service towards the trec eval package (*trec.nist.gov/trec_eval*). Trec eval was first developed and adopted by TREC and is the standard tool for computing the basic performance figures, such as precision and recall.

Finally, the 'DIRECT Service' provides the interface logic layer with uniform and integrated access to the various tools. As with the case of the 'DIRECT Datastore', the 'DIRECT Service' makes the addition of new tools transparent for the interface logic layer too.

Interface logic

The modularity of the components has enormous benefits when building interactive applications, as it helps the designer to better understand and develop each component and modify it without affecting the others. Therefore, the model-view-controller (Krasner and Pope, 1988) approach has been used 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 6.3 shows the architecture of the DIRECT user interface, which is a web-based application designed to be cross-platform and easily deployable and accessible without the need to install any software on the end-user's machine.

The system also supports the internationalisation and localisation of the user interface by adapting it to the language and country of the user. The correct language and country are initially loaded according to the browser settings and, in the case of non-supported locales, it falls back to a default configuration. The user interface has been translated into Bulgarian, Czech, English, Farsi, French, German, Indonesian, Italian, Portuguese and Spanish (Dussin and Ferro, 2007).

DIRECT: the running prototype

DIRECT has been successfully adopted in the CLEF campaigns since 2005, as reported in Table 6.1. Note that the languages of the assessed documents include not only languages with a Latin alphabet, but also Bulgarian and Russian, which use the Cyrillic alphabet, and Farsi, which is written from right to left.

Table 6.1 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 organisers, i.e. performance measures about the experiments submitted by a participant; and, lastly, knowledge produced by organisers, i.e. the statistics and statistical analyses about the different tasks of an evaluation campaign.

Table 6.1 Usage statistics of the DIRECT system

CLEF	Experiments	Participants/nations	Assessed documents	Assessors
2005	530	30/15 nations	160,000/7 languages	15
2006	570	75/25 nations	200,000/9 languages	40
2007	430	45/18 nations	215,000/7 languages	75
2008	490	40/20 nations	250,000/7 languages	65

The interface is based on a set of folding tables that allow participants to access their experiments, which are participant data, by structuring them in different levels based on a tree structure – tracks, tasks and experiments – well known to the user. The participant can therefore manage their own data by simply selecting and expanding the right level in the tree 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. By proposing only those data that are pertinent to the task currently selected by the participant, DIRECT makes available only those topics and relevance assessments which are suitable assessor data.

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

Lastly, following Zeleny (1987) who points out that knowledge is the process through which 'individual pieces of data and information (components, concepts) become connected with one another (i.e. organised) in a network of relations', the system allows users to navigate the interface and access more information resources, so that they can benefit from this 'network of relations'.

Figure 6.4 Main page for the experiment management

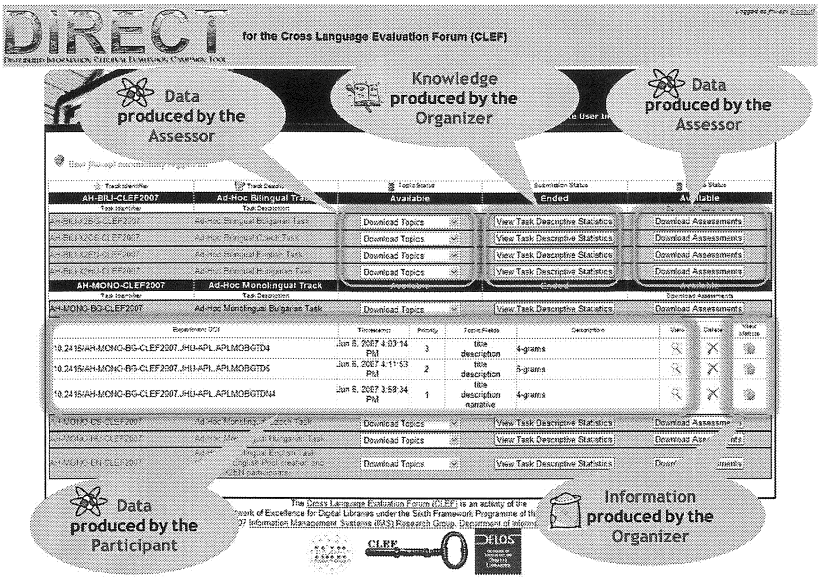
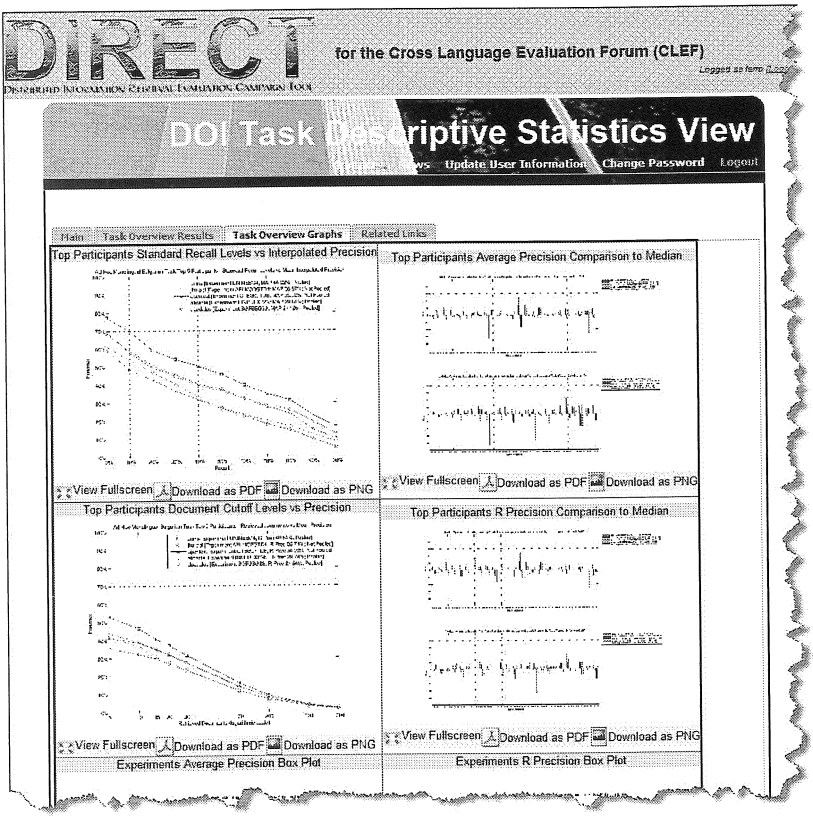


Figure 6.5 Plots about task overall performances and statistics



As an example, Figure 6.4 shows the information resources offered to the participant when the 'view tasks descriptive statistics' button is pressed. In particular, Figure 6.5 shows some of the plots used to summarise 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.

Conclusions and future work

This chapter has described the methodology currently adopted for the experimental evaluation of the information access components of a

digital library, and has shown how to extend it to include proper management, curation, archiving and enrichment of the scientific data produced while conducting an experimental evaluation in the context of large-scale evaluation campaigns.

A description has been given of the approach for maintaining the scientific output of an evaluation campaign in a DLS, in order to ensure long-term preservation, curation of data, and accessibility over time both by humans and automatic systems. The aim is to exploit the DLS for scientific data to support services for the creation, interpretation and use of multidisciplinary and multilingual digital content and to foster knowledge transfer towards relevant application communities and industry. This DLS should be constituted by several cooperating services, each focused on one aspect of digital library evaluation, where the service for managing the evaluation of the performances of the information access components represents a relevant example.

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

The introduction to this chapter underlined that the scientific reflection DLS being designed and developed must be constituted by different and cooperative services, each focused on supporting the evaluation of one aspect. As such, a DLS must supply coherent basic services and correspondent components that provide end users with access to information and documents. These services usually include an OPAC-like component, the information access components and a component to navigate the different collections of documents to which the digital library gives access. As this coherent set of components/services is of great interest to end users, these can be used to start log data collection and analysis which can be considered a parallel and correspondent effort to the one conducted during the CLEF campaigns for the collection of scientific data, described previously in this work.

The relevant characteristics of a DLS suggest addressing the issue of collecting and analysing log data in a wide and comprehensive way, otherwise many aspects of the interaction between the end user and the DLS with its diversified services may be missed (Agosti, 2008). For all the different categories of DLS users, the quality of the services and documents supplied by the digital library are very important (Agosti et al., 2007c). Log data constitute a relevant aspect in evaluating the quality of a DLS and the quality of interoperability in digital library

services. Work is underway to establish a framework for handling log data in the evaluation of DLS services that provide end users with access to information and documents, bearing in mind that the end users are central to any study of this sort. Indeed, the final user should be the guide of the system's designers, prompting them to conceive and invent solutions of real use for the user himself.

Acknowledgments

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

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

Bibliography

Abiteboul, S., Agrawal, R., Bernstein, P. A., Carey, M. J., Ceri, S., Croft, W. B., DeWitt, D. J., Franklin, M. J., Garcia-Molina, H., Gawlick, D., Gray, J., Haas, L. M., Halevy, A. Y., Hellerstein, J. M., Ioannidis, Y. E., Kersten, M. L., Pazzani, M. J., Lesk, M., Maier, D., Naughton, J. F., Schek, H.-J., Sellis, T. K., Silberschatz, A., Stonebraker, M., Snodgrass, R. T., Ullman, J. D., Weikum, G., Widom, J. and Zdonik, S. B. (2005) 'The Lowell Database research self-assessment', *Communications of the ACM* 48(5): 111–18.

Ackoff, R. L. (1989) 'From data to wisdom', *Journal of Applied Systems Analysis* 16: 3–9.

Agosti, M. (2008) 'Log data in digital libraries', in *Post-proceedings of the 4th Italian Research Conference on Digital Library Systems, Padua, 24–25 January*, Pisa: ISTI-CNR at Gruppo ALI, pp. 115–22.

Agosti, M., Di Nunzio, G. M. and Ferro, N. (2007a) 'A proposal to extend and enrich the scientific data curation of evaluation campaigns', in *Proceedings of the 1st International Workshop on*

Evaluating Information Access, Tokyo, 15 May, Tokyo: National Institute of Informatics, pp. 62–73.

Agosti, M., Di Nunzio, G. M. and Ferro, N. (2007b) 'Scientific data of an evaluation campaign: Do we properly deal with them?' In *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, Budapest, September, LNCS Vol. 4730*, Berlin, Germany: Springer-Verlag, pp. 11–20.

Agosti, M., Ferro, N., Fox, E. A. and Gonçalves, M. A. (2007c) 'Modelling DL quality – a comparison between approaches: The DELOS Reference Model and the 5S model', paper presented at the Second DELOS Conference on Digital Libraries, Tirrenia, Pisa, 5–7 December, available at: *http://www.delos.info/index.php?option=com_content&task=view&id=602&Itemid=334*.

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.

Brase, J. (2004) 'Using digital library techniques: Registration of scientific primary data', in *Proceedings of the 8th European Conference on Digital Libraries, Bath, 12–17 September, LNCS Vol. 3232*, Berlin: Springer-Verlag, pp. 488–94.

Cleverdon, C. W. (1997) 'The Cranfield tests on index languages device', in K. Spärck Jones and P. Willet (eds) *Readings in Information Retrieval*, San Francisco, CA: Morgan Kaufmann, pp. 47–60.

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*, Norwell, MA: Kluwer, pp. 1–36.

Di Nunzio, G. M., Ferro, N., Jones, G. J. and Peters, C. (2006) 'CLEF 2005: Ad hoc track overview', in *Evaluation of Multilingual and Multi-modal Information Retrieval: 6th Workshop of the Cross-Language Evaluation Forum, Alicante, 17–22 September, LNCS Vol. 4022*, Berlin: Springer-Verlag, pp. 11–36.

Dussin, M. and Ferro, N. (2007) 'Design of the user interface of a scientific digital library system for large-scale evaluation campaigns', paper presented at the Second DELOS Conference on Digital Libraries, Tirrenia, Pisa, 5–7 December, available at: *http://www.delos.info/index.php?option=com_content&task=view&id=602&Itemid=334* (accessed 15 April 2009).

Dussin, M. and Ferro, N. (2008a) 'The design of the user interface of a scientific DLS in the context of the data, information, knowledge and wisdom hierarchy', in *Post-proceedings of the 4th Italian Research*

Conference on Digital Library Systems, Padua, 24–25 January, ISTI-CNR at Gruppo ALI: Pisa, pp. 105–13.

Dussin, M. and Ferro, N. (2008b) 'Design of a digital library system for large-scale evaluation campaigns', in *Proceedings of the 12th European Conference on Digital Libraries, Aarhus, 14–19 September, LNCS Vol. 5173*, Berlin: Springer-Verlag, pp. 400–1.

Führ, N., Hansen, P., Micsik, A. and Sølvberg, I. (2001) 'Digital libraries: A generic classification scheme', in *Proceedings of the 5th European Conference on Digital Libraries, Darmstadt, 4–9 September, LNCS Vol. 2163*, Berlin: Springer-Verlag, pp. 187–99.

Führ, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S. Klas, C.-P., Kovács, L., Landoni, M., Micsik, A., Papatheodorou, C., Peters, C. and Sølvberg, I. (2007) 'Evaluation of digital libraries', *International Journal on Digital Libraries* 8(1): 21–38.

Hanks, P. (ed.) (1979) *Collins Dictionary of the English Language*, Glasgow: William Collins Sons and Co.

Harman, D. and Buckley, C. (2004) 'SIGIR 2004 Workshop: RIA and "Where can IR go from here?"', *ACM SIGIR Forum* 38(2): 45–9.

Hull, D. A. (1993) 'Using statistical testing in the evaluation of retrieval experiments', in *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA, 27 June to 1 July*, New York: ACM Press, pp. 329–38.

Ioannidis, Y. E., Maier, D., Abiteboul, S., Buneman, P., Davidson, S. B., Fox, E. A., Halevy, A. Y., Knoblock, C.A., Rabitti, F., Schek, H.-J. and Weikum, G. (2005) 'Digital library information-technology infrastructures', *International Journal on Digital Libraries* 5(4): 266–74.

Krasner, G. E. and 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.

Lord, P. and 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', available at: *http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf* (accessed 25 November 2008).

National Science Board (2005) 'Long-lived digital data collections: Enabling research and education in the 21st century', available at: *http://www.nsf.gov/pubs/2005/nsb0540/* (accessed 25 November 2008).

Paskin, N. (2005) 'Digital object identifiers for scientific data', *Data Science Journal* 4: 12–20.

Paskin, N. (ed.) (2006) 'The DOI Handbook – Edition 4.4.1', available at: *http://dx.doi.org/10.1000/186* (accessed 25 November 2008).

Tsichritzis, D. C. and Lochovsky, F. H. (1982) *Data Models*. Englewood Cliffs, NJ: Prentice Hall.

Zeleny, M. (1987) 'Management support systems: Towards integrated knowledge management', *Human Systems Management* 7(1): 59–70.

Zobel, J. (1998) 'How reliable are the results of large-scale information retrieval experiments', in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, 24–28 August*, New York: ACM Press, pp. 307–14.