A Data Curation Approach to Support In-depth Multilingual Evaluation Studies

Maristella Agosti
agosti@dei.unipd.it

Giorgio Maria Di Nunzio
dinunzio@dei.unipd.it

Nicola Ferro
ferro@dei.unipd.it

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

ABSTRACT

This work critically examines the current way of keeping the data produced during the evaluation campaigns. To overcome the shortenings of the present attitude, a new approach of considering the evaluation campaigns data as scientific data to be cured to be able to support in-depth evaluation studies is proposed and considered.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms

Design, Experimentation, Performance

Keywords

Data curation, in-depth evaluation studies

1. INTRODUCTION

The information management system evaluation turns out to be a scientific activity whose outcomes, such as performance analyses and measurements, constitute a kind of *scientific data* that need to be properly considered and used for the design and development of improved and advanced information management system components and services. By *information management system* we mean here each system able to automatically manage and retrieve information of interest for a final user, among those ones there are information retrieval systems, search engines, and digital library management systems.

When we deal with scientific data, the *provenance* (*lineage*) of the data must be tracked, and the information on the different activities of cleaning, rescaling, or modelling that have been pursued on the data must be kept, because they must be available to the researcher that makes use of the data to help him in elaborating and interpreting them [1]. When scientific data are maintained also for further and future use, they must be enriched and, together with information on provenance also information on changes at sources and on other types of changes that must have been occurred over time need to be maintained. Sometimes the enrichment of a portion of scientific data can make use of a *citation* for explicitly mentioning and making references to useful information.

Here we concentrate on a new approach to curation of scientific data that are produced by an information management system at work, and we propose a set of characteristics that this approach must have to support multilingual information access and extraction capabilities of a system of that kind. However, some of the considerations that are made can be extended also to the evaluation of other aspects and components of an information management system. So, we investigate the current evaluation methodologies adopted for assessing the performances of the information access and extraction components of a system, which deals with the indexing, search and retrieval of multilingual documents in response to a user's query.

The paper is organized as follows. Section 2 gives a critical view on the current way of keeping over time the scientific data produced during the evaluation forums. Section 3 presents a data curation approach able to support in-depth evaluation and failure studies. Section 4 closes the paper with some preliminary considerations on such a service together with some indications on future work.

2. SCIENTIFIC DATA PRODUCED IN INTERNATIONAL EVALUATION FORUMS

2.1 Scientific Data Availability over Time

Nowadays, the evaluation of the information access components of an information management system is carried out in important international evaluation forums which bring research groups together, provide them with the means for measuring the performances of their systems, discuss and compare the experimental results. The *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 systems by measuring the effectiveness of different techniques, and to discuss how differences be-

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

tween systems affected performances. After TREC, other international important initiatives have been launched, in particular *Cross-Language Evaluation Forum (CLEF)*, *NII-NACSIS Test Collection for* ***IR!*** *Systems (NTCIR)* and *INitiative for the Evaluation of XML Retrieval (INEX)*. CLEF[2] mainly aims at evaluating cross language information retrieval systems that operate on multiple languages in both monolingual and cross-lingual 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. INEX[4] provides participants with evaluation procedures for content-oriented *eXtensible Markup Language (XML)* retrieval in order to measure the effectiveness of information retrieval systems that manage XML documents. These evaluation forums are usually further organized into tracks, which investigate different facets of the evaluation of the information access components of an information management system.

2.2 Experimental Collections and Data Management

All of the previously mentioned initiatives are generally carried out according to the Cranfield methodology, which makes use of *experimental collections* [3]. An experimental collection, also named *test-collection*, is a triple $\mathcal{C} = (D, Q, J)$, where: D is a set of documents, called also collection of documents; Q is a set of topics, from which the actual queries are derived; J is a set relevance judgements, i.e. for each topic $q \in Q$ the documents $d \in D$, which are relevant for the topic q, are determined. An experimental collection $\mathcal{C}$ allows the comparison of two retrieval methods, say X and Y, according to some measurements which quantifies the retrieval performances of these methods. The most common figures adopted for quantifying the performances are the *recall*, which is a measure of the ability of a system to present all relevant items, and the *precision*, which is a measure of the ability of a system to present only relevant items. An experimental collection both provides a common test–bed to be indexed and searched by the information management systems X and Y and guarantees the possibility of replicating the experiments.

In the existing evaluation forums, this methodology is generally carried out in the following way: each participant acquires the collections of documents and the topics from the organizers of the evaluation campaign; then, he uses his own system to process the collections and topics in order to produce a list of results; finally, he returns the list of results to the organizers who compute the performance figures for his experiments.

The exchange of information between organizers and participants is mainly performed by means of textual files formatted according to the TREC data format, which is the de-facto standard in this field. As an example, the following is a fragment of the results of an experiment submitted by a participant to the organizers, where the gray header - which contains metadata that are rich of meaning for the participants to the evaluation effort - is not really present in the exchanged data but serves here as an explanation of the fields. Note that those information represent a first kind of important scientific data produced during the evaluation process, and usually those metadata are not kept and managed in an explicit way for further information and considerations.

Topic	Iter.	Document	Rank	Score	Experiment
141	Q0	AGZ.950609.0067	0	0.440873414278	IMSMIPO
141	Q0	AGZ.950613.0165	1	0.305291658641	IMSMIPO
...					

As the reader can appreciate, the fields are separated by white spaces, and they represent: the topic under evaluation, the query identifier, the unique identifier of a document, the rank of this document for the specified query, the score for the document, and the unique identifier of the experiment.

The following is an exemplary fragment of relevance judgements sent back by organizers to participants that have submitted the previous data, also in this example the gray header is not present in the exchanged data but gives metadata information on the represented fields:

Topic	Iter.	Document	Relevant
141	0	AGZ.950606.0013	0
141	0	AGZ.950609.0067	1
141	0	AGZ.950613.0165	0
...			

In the above data, each row represents a record of an experiment or of a relevance judgement, where fields are separated by white spaces. There is the field which specifies the unique identifier of the topic (e.g. `141`), the field for the unique identifier of the document (e.g. `AGZ.950609.0067`), the field which identifies the experiment (e.g. `IMSMFPO`), the field which specifies whether a document is relevant to a topic (e.g. `1`) or not (e.g. `0`).

As you can note from the above examples, this format is mainly focused on the data exchange between participants and organizers, since it allows the minimum information required to be transmitted. In the examples above, to assess the results of an experiment we need to know, at least, which documents have been retrieved in response to a given topic, or to compute the performances of a system we need the information about which documents are relevant for a given topic. On the other hand, this format is not very suitable for modelling the information space involved by an evaluation forum because, for example, the relationships among the different entities (documents, topics, experiments, participants) are not modeled and each entity is treated separately from the others.

Furthermore, present collections keeping over time does not permit systematic studies on reached improvements by participants over the years, for example in a specific multilingual setting [**?**].

2.3 Statistical Analysis of Experiments

The Cranfield methodology is mainly focused on how to evaluate the performances of two systems and how to provide a common ground which makes the experimental results comparable. [6] 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 X and Y can be considered statistically significant. To address this issue, the organizers of each evaluation forum have traditionally carried out statistical analyses, which provide participants with an overview analysis of the submitted experi-

[2] `http://clef.isti.cnr.it/`
[3] `http://research.nii.ac.jp/ntcir/index-en.html`
[4] `http://inex.is.informatik.uni-duisburg.de/`

ments, as in the case of the overview papers of the different tracks at TREC and CLEF; some recent examples of this kind of papers are [2] and [8]. Furthermore, participants may conduct statistical analyses on their own experiments by using either ad-hoc packages, such as IR-STAT-PAK[5], or general-purpose available software tools with statistical analysis capabilities, like R[6], SPSS[7], or MATLAB[8]. 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 them is not fully granted due to possible differences in the design of the statistical experiments. In fact different statistical tests can be employed to analyze the same group of results, or different choices and approximations for the various parameters of the use of the same statistical test on the same group of results can be made.

Those considerations have suggested us to re-consider the used approach in managing multilingual collections and study an approach able to overcome those limitations. The designed and adopted approach is in line with a data curation one and it is able to support in-depth and longitudinal studies over multilingual experimental collections. We have and we are still testing this new approach in managing the CLEF infrastructure in 2005 and 2006. The approach is introduced and discussed in the following section.

3. SUPPORT OF IN-DEPTH STUDIES

3.1 Data Enrichment and Interpretation

Scientific data, their enrichment and interpretation are essential components of scientific research. For the evaluation and future better development of an information retrieval component of an information management system the Cranfield methodology traces out how the relevant 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 do 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 view 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 to permit also longitudinal analyses on the data.

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 [7]: "the activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose".

This definition implies that we have to take into consideration the possibility of information enrichment of scientific data, meant as archiving and preserving scientific data so that the experiments, records, and observations will be available for future research, as well as provenance, curation, and citation of scientific data items. The benefits of this approach include the growing involvement of scientists in international research projects and forums and increased interest in comparative research activities. Furthermore, the definition introduced above reflects the importance of some of the many possible reasons for which keeping data is important, for example: re-use of data for new research, including collection based research to generate new science; retention of unique observational data which is impossible to re-create; retention of expensively generated data which is cheaper to maintain than to re-generate; enhancing existing data available for research projects; validating published research results. Not to mention the possibility of cross dissemination of scientific results with a great benefit also for industrial partners that could make use of publicly available scientific results for building new products with innovative capabilities.

As a concrete example in the field of information retrieval, please consider the data fusion problem [4], 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 even by other researchers. This is the case, for example, of the CLEF 2005 multilingual merging track [5], 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 now clear that researchers of this field would benefit by a clear data curation strategy, which promotes the reuse of existing data and allows the data experiments to be traced back to the original list of results and, perhaps, to the analyses and interpretations about them.

3.2 The Data Curation Approach

We argue that the information space implied by an evaluation forum needs an appropriate approach which takes into consideration and describes all the entities involved in the evaluation forum. In fact, an appropriate model is the necessary basis to make the scientific data produced during an evaluation campaign an active part of all those information enrichments, as data provenance and citation, we have previously described.

Thus, we can observe that, in general, there is a limited support to the systematical employment of statistical analysis by participants. For this reason, we suggest that evaluation forums should support and guide participants in adopting a more uniform way of performing statistical analyses on their own experiments. In this way, 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 make the analysis and assessment of their experiments comparable too.

The points that have been previously highlighted need to

[5] http://users.cs.dal.ca/~jamie/pubs/IRSP-overview.html
[6] http://www.r-project.org/
[7] http://www.spss.com/
[8] http://www.mathworks.com/

be considered as requirements that have to be taken into account when we are going to produce and manage scientific data that come out from the evaluation of the information access and extraction components of an information management system. In addition, to achieve the full and necessary information enrichment, both the experimental datasets and their further elaboration, such as their statistical analysis, should be first class objects that can be directly referenced and cited. Indeed, as recognized by [7], 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.

A data curation approach able to support the production and maintenance of the scientific data for in-depth evaluation studies must provide the following:

- the management of a complete evaluation forum:
  - the track set-up,
  - the harvesting of documents, and
  - 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 assessments;
- common statistical analysis tools for both organizers and participants in order to allow the comparison of experiments;
- common tools for summarizing, producing reports and graphs on the measured performances and conducted analyses;
- general metadata.

3.3 General Metadata

General metadata and metadata about experiments play an important role to keep information on the lineage of the data together with information on the programs that have to be used to process the data. Furthermore, metadata can be used to describe the context in which the different test-collections have been produced and adopted, because information on the context in which the data has been produced is essential for its study. In fact, a test-collection could have been built in a period of time in which specific social and political situations were happening and without the context knowledge of those situations, it could be impossible to make use of that test-collection in a fruitful way.

A final requirement to fulfil with the use of general metadata is that of supporting an *indefinite* electronic storage of information or a *100-year storage* as it has been defined in [1], that it means to prevent the information loss, because of storage media deterioration, or because the application to interpret the information no longer works. Keeping the metadata, data, and information necessary for managing evaluation forums in an integrated fashion in a coherent repository managed with database and digital library technologies prevent data loss giving the support for keeping the data in up-to-date formats, and to maintain information on methods that can interpret information. In particular the keeping and management of useful metadata can give the support for the fulfilment of this requirement.

4. CONCLUSION AND ONGOING WORK

The discussed data curation approach that can help to face the test-collection challenge for the evaluation and future development of information access and extraction components of interactive information management systems. On the basis of the experience gained keeping and managing the data of interest of an evaluation campaign of an international evaluation forum, we are testing the considered requirements to revise the approach and to produce an operative implementation of it.

Acknowledgements

This work 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).

5. REFERENCES

[1] S. Abiteboul et alii. The Lowell Database Research Self-Assessment. *Comm. of the ACM (CACM)*, 48(5):111–118, 2005.

[2] M. Braschler, G. M. Di Nunzio, N. Ferro, and C. Peters. CLEF 2004: Ad Hoc Track Overview and Results Analysis. In *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross–Language Evaluation Forum (CLEF 2004) Revised Selected Papers*, 10–26. LNCS 3491, Springer, Heidelberg, Germany, 2005.

[3] C. W. Cleverdon. The Cranfield Tests on Index Languages Devices. In *Readings in Information Retrieval*, 47–60. Morgan Kaufmann Publisher, Inc., San Francisco (CA), USA, 1997.

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

[5] G. M. Di Nunzio, N. Ferro, G. J. F. Jones, and C. Peters. CLEF 2005: Ad Hoc Track Overview. In *Working Notes for the CLEF 2005 Workshop*, 2005.

[6] D. Hull. Using Statistical Testing in the Evaluation of Retrieval Experiments. In *Proc. 16th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 1993)*, pages 329–338. ACM Press, New York, USA, 1993.

[7] P. Lord and A. Macdonald. *e-Science Curation Report. Data curation for e-Science in the UK*. The JISC Committee for the Support of Research (JCSR)., 2003.

[8] E. M. Voorhees. Overview of the TREC 2004 Robust Track. In *Proc. 13th Text Retrieval Conf. (TREC 2004)*. NIST, Washington, USA., 2004.