Scientific Data of an Evaluation Campaign: Do We Properly Deal with Them?

Maristella Agosti, Giorgio Maria Di Nunzio, and Nicola Ferro

Department of Information Engineering – University of Padua
Via Gradenigo, 6/b – 35131 Padova – Italy
{agosti, dinunzio, ferro}@dei.unipd.it

Abstract. This paper examines the current way of keeping the data produced during the evaluation campaigns and highlights some shortenings of it. As a consequence, we propose a new approach for improving the management evaluation campaigns' data. In this approach, the data are considered as scientific data to be cured and enriched in order to give full support to longitudinal statistical studies and long-term preservation.

1 Introduction

When we reason about the data and information produced during an evaluation campaign, we should be aware that a lot of valuable *scientific data* are produced [1,2]. Indeed, if we consider some of the outcomes of an evaluation campaign, we can see how they actually are different kinds of scientific data:

- *experiments*: the primary scientific data produced by the participants of an evaluation campaign represent the starting point for any subsequent analysis;
- *performance measurements*: metrics, such as precision and recall, are derived from the experiments and are used to evaluate the performances of an *Information Retrieval System (IRS)*
- *descriptive statistics*: the statistics, such as mean or median, are computed from the performance measurements and summarized the overall performances of an IRS or a group of IRSs;
- *statistical analyses*: different statistical techniques, such as hypothesis test, makes use of the performance measurements and the descriptive statistics in order to perform an in-depth analysis of the experiments and assess their differences.

A huge amount of the above mentioned data is produced each year during an evaluation campaign and these data are an integral part of the scientific research in the information retrieval field.

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" [3]. In addition, [4] points out how provenance is "important

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 11–20, 2007.

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, besides information about provenance, also changes at sources occurred over time need to be tracked. Sometimes the enrichment of a portion of scientific data can make use of a *citation* for explicitly mentioning and making references to useful information.

In this paper we examine whether the current methodology properly deals with the data produced during an evaluation campaign by recognizing that they are in effect valuable scientific data. Furthermore, we describe the *data curation approach* [5,6] which we have undertaken to overcome some of the shortenings of the current methodology and we have applied in designing and developing the infrastructure for the *Cross-Language Evaluation Forum (CLEF)*.

The paper is organized as follows: Section 2 introduces the motivations and the objectives of our research work; Section 3 describes the work carried out in developing the CLEF infrastructure; finally, Section 4 draws some conclusions.

2 Motivations and Objectives

2.1 Experimental Collections

Nowadays, the experimental evaluation is carried out in important international evaluation forums, such as *Text REtrieval Conference (TREC)*, CLEF, and *NII-NACSIS Test Collection for IR Systems (NTCIR)*, which bring research groups together, provide them with the means for measuring the performances of their systems, discuss and compare their work.

All of the previously mentioned initiatives are generally carried out according to the Cranfield methodology, which makes use of *experimental collections* [7]. An experimental 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. An experimental collection both provides a common test–bed to be indexed and searched by the IRS X and Y and guarantees the possibility of replicating the experiments.

Nevertheless, the Cranfield methodology is mainly focused on creating comparable experiments and evaluating the performances of an IRS rather than modeling and managing the scientific data produced during an evaluation campaign.

As an example, note that 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. Note that this information represents a first kind of scientific data produced during the

evaluation process. The following is a fragment of the results of an experiment submitted by a participant to the organizers, where the gray header is not really present in the exchanged data but serves here as an explanation of the fields.

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

In the above data, each row represents a record of an experiment, 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`), and so on, as specified by the gray headers.

As it can be noted from the above examples, this format is suitable for a simple data exchange between participants and organizers. Nevertheless, neither this format provides any metadata explaining its content nor a scheme exists in order to define the structure of each file, the data type of each field, and various constraints on the data, such as numeric floating point precision. In addition, this format does not ensure that any kind of constraint is complied with, e.g. we would avoid to retrieve the same document twice or more for the same topic. Finally, this format is not very suitable for modelling the information space involved by an evaluation forum because 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 [8].

We argue that the information space implied by an evaluation forum needs an appropriate conceptual model which takes into consideration and describes all the entities involved by the evaluation forum. 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, we have described in the previous section. This conceptual model can be also translated into an appropriate logical model in order to manage the information of an evaluation forum by using the database technology, as an example. Finally, from this conceptual model we can derive also appropriate data formats for exchanging information among organizers and participants, such as an *eXtensible Markup Language (XML)* format that complies with an XML Schema [9,10] which describes and constraints the exchanged information.

2.2 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. [11] 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, evaluation forums have traditionally carried out statistical analyses, which provide participants with an overview analysis of the submitted experiments, 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 [12,13]. Furthermore, participants may conduct statistical analyses on their own experiments by using either ad-hoc packages, such as IR-STAT-PAK[1], or generally available software tools with statistical analysis capabilities, like R[2], SPSS[3], or MATLAB[4]. 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 because, for example, 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.

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.

2.3 Information Enrichment and Interpretation

As introduced in Section 1, 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.

[1] http://users.cs.dal.ca/~jamie/pubs/IRSP-overview.html
[2] http://www.r-project.org/
[3] http://www.spss.com/
[4] http://www.mathworks.com/

Let us consider what is currently done in an evaluation forum:

- Experimental collections:
  - there are few or no metadata about document collections, the context they refer to, how they have been created, and so on;
  - there are few or no metadata about topics, how they have been created, the problems encountered by their creators, what documents creators found relevant for a given topic, and so on;
  - there are few or no metadata about how pools have been created and about the relevance assessments, the problems which have been faced by the assessors when dealing with difficult topics;
- Experiments:
  - there are few or no metadata about them, such as what techniques have been adopted or what tunings have been carried out;
  - they can be not publicly accessible, making it difficult for other researchers to make a direct comparison with their own experiments;
  - their citation can be an issue;
- Performance measurements:
  - there are no metadata about how a measure has been created, which software has been used to compute it, and so on;
  - often only summaries are publicly available while all the detailed measurements may be not accessible;
  - their format can be not suitable for further computer processing;
  - their modelling and management needs to be dealt with;
- Descriptive statistics and hypothesis tests:
  - they are mainly limited to task overviews produced by organizers;
  - participants may not have all the skills needed to perform a statistical analysis;
  - participants can carry out statistical analyses only on their own experiments without the possibility of comparing them with the experiments of other participants;
  - the comparability among the statistical analyses conducted by the participants is not fully granted due to possible differences in the design of the statistical experiments.

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

This definition implies that we have to take into consideration the possibility of information enrichment of scientific data, meant as archiving and preserving scientific data so that the experiments, records, and observations will be available for future research, as well as provenance, curation, and citation of scientific data

items. The benefits of this approach include the growing involvement of scientists in international research projects and forums and increased interest in comparative research activities. Furthermore, the definition introduced above reflects the importance of some of the many possible reasons for which keeping data is important, for example: re-use of data for new research, including collection based research to generate new science; retention of unique observational data which is impossible to re-create; retention of expensively generated data which is cheaper to maintain than to re-generate; enhancing existing data available for research projects; validating published research results.

As a concrete example in the field of information retrieval, please consider the data fusion problem [15], 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 [12], 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 re-use of existing data and allows the data fusion experiments to be traced back to the originary list of results and, perhaps, to the analyses and interpretations about them.

Thus, we consider all these points as requirements that should be taken into account when we are going to produce and manage scientific data that come out from the experimental evaluation of an IRS. In addition, to achieve the full information enrichment discussed in Section 1, 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 [14], 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.

3 The CLEF Infrastructure

3.1 Conceptual Model for an Evaluation Forum

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

Figure 1 shows the *Entity–Relationship (ER)* schema which represents the conceptual model we have developed. The conceptual model is built around five main areas of modelling:

- **evaluation forum**: 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;

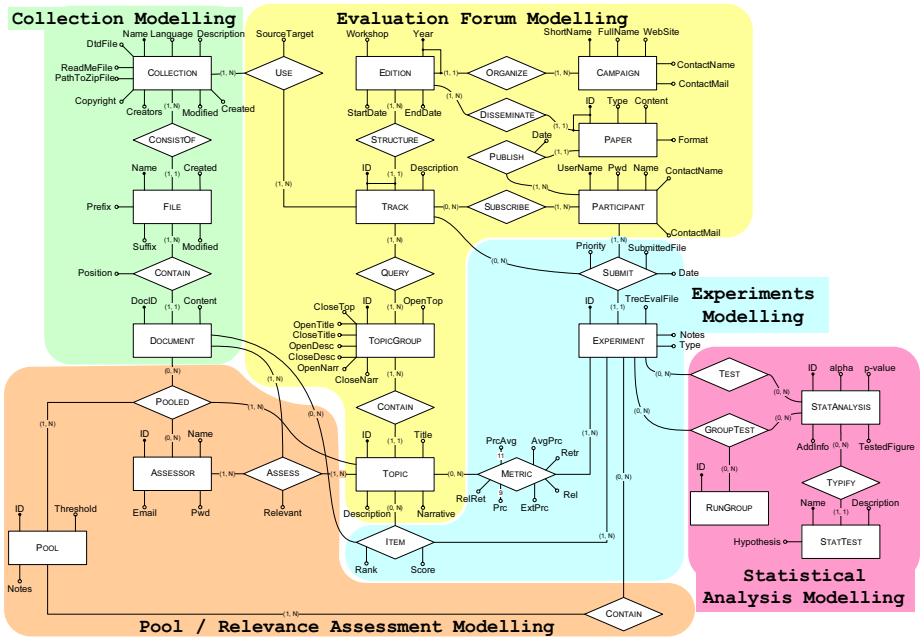

Fig. 1. Conceptual model for the information space of an evaluation forum

- **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 [16], 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.

3.2 DIRECT: The Running Prototype

The proposed approach has been implemented in a prototype, called *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* [17,18,19], and it has been tested in the context of the CLEF 2005 and 2006 evaluation campaigns. The initial prototype moves a first step towards a data curation approach to evaluation initiatives, by providing 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;
- common tools for summarizing, producing reports and graphs on the measured performances and conducted analyses;
- common 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 [20,21].

DIRECT has been developed by using the Java[5] programming language, which ensures great portability of the system across different platforms. We used the PostgreSQL[6] *DataBase Management System (DBMS)* for performing the actual storage of the data. Finally, all kinds of *User Interface (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[7] framework, an open-source framework for developing Web applications.

4 Conclusions

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 the CLEF evaluation campaign, we are testing the considered requirements to revise the proposed approach.

A prototype of the carrying out the proposed approach, called DIRECT, has been implemented and widely tested during the 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 implement a second version of the prototype to be tested and validated during the next CLEF campaigns.

[5] `http://java.sun.com/`
[6] `http://www.postgresql.org/`
[7] `http://struts.apache.org/`

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).

References

1. Agosti, M., Di Nunzio, G.M., Ferro, N.: The Importance of Scientific Data Curation for Evaluation Campaigns. In: DELOS Conference 2007 Working Notes, ISTI-CNR, Gruppo ALI, Pisa, Italy, pp. 185–193 (2007)
2. Agosti, M., Di Nunzio, G.M., Ferro, N.: A Proposal to Extend and Enrich the Scientific Data Curation of Evaluation Campaigns. In: Proc. 1st International Workshop on Evaluating Information Access (EVIA 2007), National Institute of Informatics, Tokyo, Japan, pp. 62–73 (2007)
3. Abiteboul, S., Agrawal, R., Bernstein, P., Carey, M., Ceri, S., Croft, B., DeWitt, D., Franklin, M., Garcia-Molina, H., Gawlick, D., Gray, J., Haas, L., Halevy, A., Hellerstein, J., Ioannidis, Y., Kersten, M., Pazzani, M., Lesk, M., Maier, D., Naughton, J., Schek, H.J., Sellis, T., Silberschatz, A., Stonebraker, M., Snodgrass, R., Ullman, J.D., Weikum, G., Widom, J., Zdonik, S.: The Lowell Database Research Self-Assessment. Communications of the ACM (CACM) 48, 111–118 (2005)
4. Ioannidis, Y., Maier, D., Abiteboul, S., Buneman, P., Davidson, S., Fox, E.A., Halevy, A., Knoblock, C., Rabitti, F., Schek, H.J., Weikum, G.: Digital library information-technology infrastructures. International Journal on Digital Libraries 5, 266–274 (2005)
5. Agosti, M., Di Nunzio, G.M., Ferro, N.: A Data Curation Approach to Support In-depth Evaluation Studies. In: Proc. International Workshop on New Directions in Multilingual Information Access (MLIA 2006), pp. 65–68 (2006) (last visited, March 23, 2007), http://ucdata.berkeley.edu/sigir2006-mlia.htm
6. Agosti, M., Di Nunzio, G.M., Ferro, N., Peters, C.: CLEF: Ongoing Activities and Plans for the Future. In: Proc. 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, National Institute of Informatics, Tokyo, Japan, pp. 493–504 (2007)
7. Cleverdon, C.W.: The Cranfield Tests on Index Languages Devices. In: Readings in Information Retrieval, pp. 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, California, USA (1997)
8. Agosti, M., Di Nunzio, G.M., Ferro, N.: Evaluation of a Digital Library System. In: Notes of the DELOS WP7 Workshop on the Evaluation of Digital Libraries, pp. 73–78 (2004) (last visited, March 23, 2007), http://dlib.ionio.gr/wp7/workshop2004_program.html
9. W3C: XML Schema Part 1: Structures - W3C Recommendation 28 October 2004. (2004) (last visited, March 23, 2007), http://www.w3.org/TR/xmlschema-1/
10. W3C: XML Schema Part 2: Datatypes - W3C Recommendation 28 October 2004. (2004) (last visited, March 23, 2007), http://www.w3.org/TR/xmlschema-2/
11. Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments. In: Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993), pp. 329–338. ACM Press, New York, USA (1993)

12. Di Nunzio, G.M., Ferro, N., Jones, G.J.F., Peters, C.: CLEF 2005: Ad Hoc Track Overview. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 11–36. Springer, Heidelberg (2006)
13. Voorhees, E.M.: Overview of the TREC 2005 Robust Retrieval Track. In: The Fourteenth Text REtrieval Conference Proceedings (TREC 2005), National Institute of Standards and Technology (NIST), Special Pubblication 500-266, Washington, USA (2005) (last visited, March 23, 2007), http://trec.nist.gov/pubs/trec14/t14_proceedings.html
14. Lord, P., Macdonald, A.: 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) (2003) (last visited, March 23, 2007), http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf
15. Croft, W.B.: Combining Approaches to Information Retrieval. In: Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval, pp. 1–36. Kluwer Academic Publishers, Norwell (MA), USA (2000)
16. Harman, D.K.: Overview of the First Text REtrieval Conference (TREC-1). In The First Text REtrieval Conference (TREC-1), National Institute of Standards and Technology (NIST), Special Pubblication 500-207, Washington, USA (1992) (last visited, March 23, 2007), http://trec.nist.gov/pubs/trec1/papers/01.txt
17. Di Nunzio, G.M., Ferro, N.: DIRECT: a Distributed Tool for Information Retrieval Evaluation Campaigns. In: Proc. 8th DELOS Thematic Workshop on Future Digital Library Management Systems: System Architecture and Information Access, pp. 58–63 (2005) (last visited, March 23, 2007), http://dbis.cs.unibas.ch/delos_website/delos-dagstuhl-handout-all.pdf
18. Di Nunzio, G.M., Ferro, N.: DIRECT: a System for Evaluating Information Access Components of Digital Libraries. In: Rauber, A., Christodoulakis, S., Tjoa, A.M. (eds.) ECDL 2005. LNCS, vol. 3652, pp. 483–484. Springer, Heidelberg (2005)
19. Di Nunzio, G.M., Ferro, N.: Scientific Evaluation of a DLMS: a service for evaluating information access components. In: Gonzalo, J., Thanos, C., Verdejo, M.F., Carrasco, R.C. (eds.) ECDL 2006. LNCS, vol. 4172, pp. 536–539. Springer, Heidelberg (2006)
20. Di Nunzio, G.M., Ferro, N.: Appendix A. Results of the Core Tracks and Domain-Specific Tracks. In: Peters, C., Quochi, V. (eds.) Working Notes for the CLEF 2005 Workshop (2005) (last visited, March 23, 2007), http://www.clef-campaign.org/2005/working_notes/workingnotes2005/appen dix_a.pdf
21. Di Nunzio, G.M., Ferro, N.: Appendix A: Results of the Ad-hoc Bilingual and Monolingual Tasks. In: Nardi, A., Peters, C., Vicedo, J.L. (eds.) Working Notes for the CLEF 2006 Workshop (2006) (last visited, March 23, 2007), http://www.clefcampaign.org/2006/working_notes/workingnotes2006/Appendix_Ad-Hoc.pdf