The Importance of Scientific Data Curation for Evaluation Campaigns

Maristella Agosti, Giorgio Maria Di Nunzio, and Nicola Ferro

Department of Information Engineering, University of Padua, Italy
{agosti,dinunzio,ferro}@dei.unipd.it

Abstract. Information Retrieval system evaluation campaigns produce valuable scientific data, which should be preserved carefully so that they can be available for further studies. A complete record should be maintained of all analyses and interpretations in order to ensure that they are reusable in attempts to replicate particular results or in new research and so that they can be referred to or cited at any time.

In this paper, we describe the data curation approach for the scientific data produced by evaluation campaigns. The medium/long-term aim is to create a large-scale *Digital Library System (DLS)* of scientific data which supports services for the creation, interpretation and use of multidisciplinary and multilingual digital content.

1 Introduction

The experimental evaluation of *Information Retrieval (IR)* systems is usually carried out in important international evaluation campaigns, such as *Text REtrieval Conference (TREC)*[1], *Cross-Language Evaluation Forum (CLEF)*[2], and *NII-NACSIS Test Collection for IR Systems (NTCIR)*[3], which bring research groups together, providing them with the means to compare the performances of their systems, and discuss their results. This paper examines the approach traditionally adopted for experimental evaluation in the IR research field in the light of the challenges posed by the recent recognition of the importance of a correct management, preservation and access to scientific data. We discuss how the increasing attention being given to this question impacts on both IR evaluation methodology and on the way in which the data of the evaluation campaigns are organized and maintained over time.

The paper is organized as follows: Section 2 introduces the motivations and the objectives of our research work. Section 3 discusses possible ways of extending the current evaluation methodology both from the point of view of the the conceptual model of the information space involved and a software infrastructure for evaluation campaigns. Section 4 draws some conclusions.

[1] http://trec.nist.gov/
[2] http://www.clef-campaign.org/
[3] http://research.nii.ac.jp/ntcir/index-en.html

C. Thanos, F. Borri, and L. Candela (Eds.): Digital Libraries: R&D, LNCS 4877, pp. 157–166, 2007.

2 Evaluation Campaigns and IR Experimental Evaluation

Much IR system evaluation is based on a comparative evaluation approach in which system performances are compared according to the Cranfield methodology, which makes use of test collections [1]. A test collection $\mathcal{C}$ allows the comparison of information access systems according to measurements which quantify their performances. The main goals of a test collection are to provide a common test-bed to be indexed and searched by information access systems and to guarantee the possibility of replicating the experiments.

2.1 Methodology

If we consider the Cranfield evaluation methodology and the achievements and outcomes of the evaluation campaigns in which it is used, it is clear that we are dealing with different kinds of valuable *scientific data*. The test collections and the experiments represent our primary scientific data and the starting point of our investigation. Using the test data, we produce different performance measurements, such as precision and recall, in order to evaluate the performances of IR systems for a given experiment. Starting from these performance measurements, we compute descriptive statistics, such as mean or median, which can be used to summarize the overall performances achieved by an experiment or by a collection of experiments. Finally, we perform hypothesis tests and other statistical analyses in order to conduct in-depth studies and comparisons over a set of experiments.

We can frame the above mentioned scientific data in the context of the *Data, Information, Knowledge, Wisdom (DIKW)* hierarchy [2,3]:

- *data*: the *test collections* and the *experiments* correspond to the "data level" in the hierarchy, since they are the raw, basic elements needed for any further investigation and have little meaning by themselves. An experiment and the results obtained by conducting it are almost useless without knowledge of the test collection used for the experiment; these data constitute the basis for any subsequent computation;
- *information*: the *performance measurements* correspond to the "information level" in the hierarchy, since they are the result of computations and processing on the data; in this way, we can give meaning to the data via certain relations. For example, precision and recall measures are obtained by relating the list of results contained in an experiment with the relevance judgements J;
- *knowledge*: the *descriptive statistics* and the *hypothesis tests* correspond to the "knowledge level" in the hierarchy, since they represent further processing of the information provided by the performance measurements and provide us with some insights about the experiments;
- *wisdom*: *theories*, *models*, *algorithms*, *techniques*, and *observations*, which are usually communicated by means of papers, talks, and seminars, correspond to the "wisdom level" in the hierarchy, since they provide interpretation, explanation, and formalization of the content of the previous levels.

As observed by [3], "while data and information (being components) can be generated per se, i.e. without direct human interpretation, knowledge and wisdom (being relations) cannot: they are human- and context-dependent and cannot be contemplated without involving *human* (not machine) comparison, decision making and judgement". This observation also fits the case of IR system experimental evaluation. In fact, experiments (data) and performance measurements (information) are usually generated automatically by programs, and tools for performance assessment. However, statistical analyses (knowledge) and models and algorithms (wisdom) require a deep involvement of researchers in order to be conducted and developed.

This view of IR system experimental evaluation raises the question of whether the Cranfield methodology is able to support an approach where the whole process from data to wisdom is taken into account. This question is made more compelling by the fact that, when we deal with scientific data, "the lineage (provenance) of the data must be tracked, since a scientist needs to know where the data came from [...] and what cleaning, rescaling, or modelling was done to arrive at the data to be interpreted" [4]. Moreover, as pointed out by [5], 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 are liable to be enriched and, sometimes, the enrichment of a portion of scientific data implies a *citation* so that useful information can be explicitly mentioned and referenced [6,7]. Finally, [8] highlights that "digital data collections enable analysis at unprecedented levels of accuracy and sophistication and provide novel insights through innovative information integration". Thus, the question is not only to which degree the Cranfield methodology supports passing from data to wisdom, but also whether correct strategies are adopted to ensure the provenance, the enrichment, the citation, and the interpretation of the scientific data.

2.2 Infrastructure

There is a growing interest in the proper management of scientific data by diverse world organizations, among them the European Commission (EC), the US National Scientific Board, and the Australian Working Group on Data for Science. The EC in the i2010 Digital Library Initiative clearly states that "digital repositories of scientific information are essential elements to build European eInfrastructure for knowledge sharing and transfer, feeding the cycles of scientific research and innovation up-take" [9]. The US National Scientific Board points out that "organizations make choices on behalf of the current and future user community on issues such as collection access; collection structure; technical standards and processes for data curation; ontology development; annotation; and peer review". Moreover, those organizations "are uniquely positioned to take leadership roles in developing a comprehensive strategy for long-lived digital data collections" [8]. The Australian Working Group on Data for Science suggests to establishing "a nationally supported long-term strategic framework

for scientific data management, including guiding principles, policies, best practices and infrastructure", that "standards and standards-based technologies be adopted and that their use be widely promoted to ensure interoperability between data, metadata, and data management systems", and that "the principle of open equitable access to publicly-funded scientific data be adopted wherever possible [...] As part of this strategy, and to enable current and future data and information resources to be shared, mechanisms to enable the discovery of, and access to, data and information resources must be encouraged" [10].

These observations suggest that considering the IR system experimental evaluation as a source of scientific data entails not only re-thinking the evaluation methodology itself, but also re-considering the way in which this methodology is applied and how the evaluation campaigns are organized. Indeed, changes to IR system evaluation methodology need to be correctly supported by organizational, hardware, and software infrastructure which allow for the management, search, access, curation, enrichment, and citation of the scientific data produced.

Such changes will also impact on the organizations which run the evaluation campaigns, since they have not only to provide the infrastructure but also to participate in the design and development of it. In fact, as highlighted by [8], they should take a leadership role in developing a comprehensive strategy for the preservation of digital data collections and drive the research community through this process in order to improve the way of doing research. As a consequence, the aim and the reach of an evaluation campaign would be widened because, besides bringing research groups together and providing them with the means for discussing and comparing their work, an evaluation campaign should also take care of defining guiding principles, policies, best practices for making use of the scientific data produced during the evaluation campaign itself.

3 Extending the Approach to IR Evaluation

As observed in the previous section, scientific data, their curation, enrichment, and interpretation are essential components of scientific research. These issues are better faced and framed in the wider context of the *curation of scientific data*, which plays an important role in the systematic definition of an appropriate methodology to manage and promote the use of data. The e-Science Data Curation Report gives the following definition of data curation [11]: "the activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purposes, 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; this means that we must archive and preserve scientific data so that experiments, records, and observations will be available for future research, together with information on 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.

There are many reasons why the preservation of the data resulting from an evaluation campaign is important, for example: the 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; enhancement of existing data available for research projects; validation of published research results. However, it should be remembered that the Cranfield methodology was developed to create comparable experiments and evaluate the performances of IR systems rather than to model, manage, and curate the scientific data produced during an evaluation campaign.

In the following sections, we discuss some key points to be taken into consideration when extending the current evaluation methodology in order to give evaluation campaigns a leadership role in driving research on IR evaluation methodologies.

3.1 Conceptual Model and Metadata

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

Moreover, [12] points out that "metadata descriptions are as important as the data values in providing meaning to the data, and thereby enabling sharing and potential future useful access". Since there is no conceptual model for an experimental collection, metadata schemes for describing it are also lacking. Consider that there are almost no metadata:

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

The situation is a little bit less problematic when it comes to experiments for which some kind of metadata may be collected, such as which topic fields were used to create the query, whether the query was automatically or manually constructed from the topics and, in some tracks of TREC, some information about the hardware used to run the experiments. Nevertheless, a better description of the experiments could be achieved if we take into consideration what retrieval model was applied, what algorithms and techniques were adopted, what kind of stop word removal and/or stemming was performed, what tunings were carried out.

A good attempt in this direction is the *Reliable Information Access (RIA)* Workshop [13], organized by the US *National Institute of Standards and Technology (NIST)* in 2003, where an in-depth study and failure analysis of the conducted experiments were performed and valuable information about them was collected. However, the existence of a commonly agreed conceptual model and metadata schemas would have helped in defining and gathering the information to be kept.

Similar considerations hold also for the performance measurements, the descriptive statistics, and the statistical analyses which are not explicitly modeled and for which no metadata schema is defined. It would be useful to define at least the metadata that are necessary to describe which software and which version of the software were used to compute a performance measure, which relevance judgements were used to compute a performance measure, and when the performance measure was computed. Similar metadata could be useful also for descriptive statistics and statistical analyses.

3.2 Unique Identification Mechanism

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

The absence of a mechanism to uniquely identify and reference the digital objects of an evaluation campaign prevent us from directly citing those digital objects. Indeed, as recognized by [11], the possibility of citing scientific data and their further elaboration 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, discussed in Section 2, because experimental collections and experiments would become citable and accessible just like any other item in the reference list of a paper.

Over the past years, various syntaxes, mechanisms, and systems have been developed to provide unique identifiers for digital objects, among them the following are candidates to be adopted in the unique identification of the different digital objects involved in an evaluation campaign: *Uniform Resource Identifier (URI)* [14], *Digital Object Identifier (DOI)* [15], OpenURL [16], and

Persistent URL (PURL)[4]. An important aspect of all the identification mechanisms described above is that all of them provide facilities for resolving the identifiers. This means that all those mechanisms enable a 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.

The DOI constitutes a valuable possibility for identifying and referencing the digital objects of an evaluation campaign, since there have already been successful attempts to apply it to scientific data and it makes possible the association of metadata with the identified digital objects [17,18].

3.3 Statistical Analyses

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

To address this issue, evaluation campaigns have traditionally supported and carried out statistical analyses, which provide participants with an overview analysis of the submitted experiments. Furthermore, participants may conduct statistical analyses on their own experiments by using either ad-hoc packages, such as IR-STAT-PAK[5], or generally available software tools with statistical analysis capabilities, like R[6], SPSS[7], or MATLAB[8]. However, the choice of whether to perform a statical analysis or not is left up to each participant who may even not have all the skills and resources needed to perform such analyses. Moreover, when participants perform statistical analyses using their own tools, the comparability among these analyses could not be fully granted, in fact, different statistical tests can be employed to analyze the data, or different choices and approximations for the various parameters of the same statistical test can be made.

In developing an infrastructure for improving the support given to participants by an evaluation campaign, it could be advisable to add some form of support and guide to participants for adopting a more uniform way of performing statistical analyses on their own experiments. If this support is added, participants can not only benefit from standard experimental collections which make their experiments comparable, but they can also exploit standard tools for the analysis of the experimental results, which would make the analysis and assessment of their experiments comparable too.

As recalled in Section 2, scientific data, their enrichment and interpretation are essential components of scientific research. The Cranfield methodology traces

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

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 methodologies do not require any particular coordination or synchronization between the basic scientific data and the analyses on them, which are treated as almost separate items. However, 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.

3.4 IR Evaluation and Digital Library Systems

We consider all the abovementioned key points as requirements that should be taken into account when designing an evaluation campaign which will take on a twofold role. Firstly, an evaluation campaign aims at promoting research in the IR field by highlighting valuable areas which need to be explored and by offering the means for conducting, comparing, and discussing experiments. Secondly, an evaluation campaign has to make the management and the curation of the produced scientific data an integral part of the IR research process. Therefore, an evaluation campaign has to provide guidelines, best practices, conceptual and logical models for data representation and exchange, preservation and curation of the produced scientific data, and support for passing through the whole DIKW hierarchy.

As a consequence, an evaluation campaign has to provide a software infrastructure suitable for carrying out this second new role. In this context, DLSs are the natural choice for managing, making accessible, citing, curating, enriching, and preserving all the information resources produced during an evaluation campaign. Indeed, [5] points out how *information enrichment* should be one of the activities supported by a DLS and, among the different kinds of them, considers provenance as "important in judging the quality and applicability of information for a given use and for determining when changes at sources require revising derived information". In addition, [5] observes that *citation*, intended as the possibility of explicitly mentioning and making references to portions of a given digital object, should also be part of the information enrichment strategies supported by a DLS.

In addition, the evaluation of complex systems, such as a DLS, is a non trivial issue which should analyze different aspects, among which: architecture, information access and extraction capabilities, management of multimedia content, interaction with users, and so on [20]. Since there are so many aspects to take into consideration, a DLS, which is used as an infrastructure for an evaluation campaign, should be constituted by different and cooperating services, each one focused on supporting the evaluation of one of the aspects mentioned above. This approach to the design of such DLSs is coherent with the guidelines proposed in [5], who traces out the service-based design as one of the key points in the scientific development of DLSs.

In conclusion, DLSs can act as the systems of choice to support evaluation campaigns in making a step forward; they are able to both address the key points highlighted above and provide a more mature way of dealing with the scientific data produced during the IR experimental evaluation.

4 Conclusion

This study has addressed the methodology currently adopted for the experimental evaluation in the IR field, and it has proposed extending it to include a proper management, curation, archiving, and enrichment of the scientific data that are produced while conducting an experimental evaluation in the context of evaluation campaigns. We described the approach for maintaining in a DLS the scientific output of an evaluation campaign, in order to ensure long-term preservation, curation of data, and accessibility over time both by humans and automatic systems. The aim is to create a large-scale *Digital Library (DL)* of scientific data which supports services for the creation, interpretation and use of multidisciplinary and multilingual digital content.

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. 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)
2. Ackoff, R.L.: From Data to Wisdom. Journal of Applied Systems Analysis 16, 3–9 (1989)
3. Zeleny, M.: Management Support Systems: Towards Integrated Knowledge Management. Human Systems Management 7, 59–70 (1987)
4. Abiteboul, S., et al.: The Lowell Database Research Self-Assessment. Communications of the ACM (CACM) 48, 111–118 (2005)
5. Ioannidis, Y., et al.: Digital library information-technology infrastructures. International Journal on Digital Libraries 5, 266–274 (2005)
6. Agosti, M., Di Nunzio, G.M., Ferro, N.: A Data Curation Approach to Support In-depth Evaluation Studies. In: MLIA 2006. Proc. International Workshop on New Directions in Multilingual Information Access, pp. 65–68, [last visited 2007, March 23] (2006), `http://ucdata.berkeley.edu/sigir2006-mlia.htm`
7. Agosti, M., Di Nunzio, G.M., Ferro, N.: Scientific Data of an Evaluation Campaign: Do We Properly Deal With Them? In: CLEF 2006. LNCS, vol. 4730, pp. 11–20. Springer, Heidelberg (2007)

8. National Science Board: Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century (NSB-05-40). National Science Foundation (NSF). [last visited 2007, March 23] (2005), http://www.nsf.gov/pubs/2005/nsb0540/
9. European Commission Information Society and Media: i2010: Digital Libraries. [last visited 2007, March 23] (2006), http://europa.eu.int/information_society/activities/digital_libraries/doc/brochures/dl_brochure_2006.pdf
10. Working Group on Data for Science: FROM DATA TO WISDOM: Pathways to Successful Data Management for Australian Science. Report to Minister's Science, Engineering and Innovation Council (PMSEIC), [last visited 2007, March 23] (2006), http://www.dest.gov.au/sectors/science_innovation/publications_resources/profiles/Presentation_Data_for_Science.htm
11. 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). [last visited 2007, March 23] (2003), http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf
12. Anderson, W.L.: 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 (2004)
13. Harman, D., Buckley, C.: The NRRC Reliable Information Access (RIA) Workshop. In: SIGIR 2004. Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 528–529. ACM Press, New York (2004)
14. Berners-Lee, T., Fielding, R., Irvine, U.C., Masinter, L.: Uniform Resource Identifiers (URI): Generic Syntax. RFC 2396 (1998)
15. Paskin, N., (ed.): The DOI Handbook – Edition 4.4.1. International DOI Foundation (IDF). [last visited 2007, August 30] (2006), http://dx.doi.org/10.1000/186
16. NISO: ANSI/NISO Z39.88 - 2004 – The OpenURL Framework for Context-Sensitive Services. National Information Standards Organization (NISO). [last visited 2007, March 23] (2005), http://www.niso.org/standards/standard_detail.cfm?std_id=783
17. Brase, J.: Using Digital Library Techniques – Registration of Scientific Primary Data. In: Heery, R., Lyon, L. (eds.) ECDL 2004. LNCS, vol. 3232, pp. 488–494. Springer, Heidelberg (2004)
18. Paskin, N.: Digital Object Identifiers for Scientific Data. Data Science Journal 4, 12–20 (2005)
19. Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments. In: SIGIR 1993. Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 329–338. ACM Press, New York (1993)
20. Fuhr, N., Hansen, P., Micsik, A., Sølvberg, I.: Digital Libraries: A Generic Classification Scheme. In: Constantopoulos, P., Sølvberg, I.T. (eds.) ECDL 2001. LNCS, vol. 2163, pp. 187–199. Springer, Heidelberg (2001)