DIRECTions: Design and Specification of an IR Evaluation Infrastructure

Maristella Agosti, Emanuele Di Buccio, Nicola Ferro, Ivano Masiero, Simone Peruzzo, and Gianmaria Silvello

Department of Information Engineering, University of Padua, Italy
{agosti,dibuccio,ferro,masieroi,peruzzos,silvello}@dei.unipd.it

Abstract. *Information Retrieval (IR)* experimental evaluation is an essential part of the research on and development of information access methods and tools. Shared data sets and evaluation scenarios allow for comparing methods and systems, understanding their behaviour, and tracking performances and progress over the time. On the other hand, experimental evaluation is an expensive activity in terms of human effort, time, and costs required to carry it out.

Software and hardware infrastructures that support experimental evaluation operation as well as management, enrichment, and exploitation of the produced scientific data provide a key contribution in reducing such effort and costs and carrying out systematic and throughout analysis and comparison of systems and methods, overall acting as enablers of scientific and technical advancement in the field. This paper describes the specification for an IR evaluation infrastructure by conceptually modeling the entities involved in IR experimental evaluation and their relationships and by defining the architecture of the proposed evaluation infrastructure and the APIs for accessing it.

1 Motivations

IR has always been a scientific field strongly rooted in experimentation and collaborative evaluation efforts [1]. Large-scale evaluation initiatives, such as TREC in the United States[1], the CLEF in Europe[2], and the NTCIR in Asia[3], contribute significantly to advancements in research and industrial innovation in the IR field, and to the building of strong research communities. Beside their scientific and industrial impact, a recent study conducted by NIST highlighted also the economic impact and value of large-scale evaluation campaigns and reported that for every $1 that NIST and its partners invested in TREC, at least $3.35 to $5.07 in benefits accrued to IR researchers and developers while the overall investment in TREC has been estimated in about 30 million dollars [2].

IR evaluation is challenged by variety and fragmentation in many respects – diverse tasks and metrics, heterogeneous collections, different systems, and alternative approaches for managing the experimental data. Not only does this

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

T. Catarci et al. (Eds.): CLEF 2012, LNCS 7488, pp. 88–99, 2012.

hamper the generalizability and exploitability of the results but it also increases the effort and costs needed to produce such experimental results and to further exploit them. Abstracting over these constituents as well as over the obtained results is crucial for scaling-up evaluation and evaluation infrastructures are a fundamental part of this wider abstraction process [3].

When it comes to analysis of the experimental results, several methodologies, metrics, and statistical techniques have been proposed over the years [4]. Nevertheless, it is often difficult to apply them properly, sometimes due to their complexity and the required competencies, and in a way that eases further comparison and interpretations, e.g. by choosing similar values for parameters or equivalent normalization and transformation strategies for the data. An evaluation infrastructure should not only facilitate this day-to-day analyses but it should also be able to preserve and provide access over time to them in order to support and foster the conduction and automation of longitudinal studies which track the evolution of the performances over the time, as for example [5].

An important, but often overlooked, part of analysis is the visualization of the experimental data. Visualization in IR has been mostly applied to alternative presentation of search results [6,7] while, when applied to the experimental data, it can greatly impact their comprehension and understanding, as it has been done in [8,9]. Not only visualization but also visual interaction with the experimental data and analytical models supporting such interaction, as those developed in the *Visual Analytics (VA)* field [10], should be exploited in order to facilitate the exploration and study of the experimental data. This latter possibility, i.e. VA for IR evaluation, becomes feasible only when there is an evaluation infrastructure supporting and implementing such analytical and interaction models and it has been started to be studied only very recently [11,12].

This paper presents the specification of the evaluation infrastructure which is being developed in the context of the PROMISE project[4], an European network of excellence which aims at improving the access to the scientific data produced during evaluation activities, supporting the organization and running of evaluation campaigns, increasing automation in the evaluation process, and fostering the usage of the managed scientific data. Three key contributions are discussed: (i) the conceptual schema which describes the entities involved in the experimental evaluation and the relationships among them; (ii) the architecture of the evaluation infrastructure able to manage scientific data; (iii) a set of Web API to interact with all the resources managed by the system.

The presentation is organized as follows: Section 2 reports on previous works; Section 3 describes the results of the conceptual modeling; Section 4 gives an overview of the architecture; and Section 5 presents a use case scenario based on the visualization of topics, experiments, and metrics; finally, Section 6 draws some final remarks.

[4] `http://www.promise-noe.eu/`

2 Background

During their life-span, large-scale evaluation campaigns produce valuable amounts of scientific data which are the basis for the subsequent scientific work and system development, thus constituting an essential reference for the field. Until a few years ago, limited attention had been paid to the modeling, management, curation, and access of the produced scientific data, even though the importance of scientific data in general has been highlighted by many different institutional organizations, such as the European Commission [13].

The research group on Information Management Systems of the Department of Information Engineering of the University of Padua[5] started a few years ago the challenge of addressing the most common limitations on facing the issue [14] and working on envisaging and defining a necessary infrastructure for dealing with the complexity of the challenge. We have proposed an extension to the traditional evaluation methodology in order to explicitly take into consideration and model the valuable scientific data produced during an evaluation campaign, the creation of which is often expensive and not easily reproducible. Indeed, researchers not only benefit from having comparable experiments and a reliable assessment of their performances, but they also take advantage of the possibility of having an integrated vision of the scientific data produced, together with their analyses and interpretations, as well as benefiting from the possibility of keeping, re-using, preserving, and curating them. Moreover, the way in which experimental results are managed is an integral part of the process of knowledge transfer and sharing towards relevant application communities, which needs to properly understand these experimental results in order to create and assess their own systems. Therefore, we have undertaken the design of an evaluation infrastructure for large-scale evaluation campaigns and the outcome is the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)*, which manages the scientific data produced during a large-scale evaluation campaign, as well as supports the archiving, access, citation, dissemination, and sharing of the experimental results [15,16].

In the context of the international DESIRE workshop [17], the necessity for open and public benchmarks and infrastructures has been confirmed as they represent the foundations of the scientific method adopted in the IR community. Algorithms and solutions tested and evaluated on private data not publicly accessible make it difficult for researchers and developers to reproduce them, verify their performances, and compare them with the state-of-the-art or with own solutions. Another important point that has been highlighted is the need for a proper and shared modeling of the experimental data produced by IR evaluation, in terms of conceptual model, descriptive metadata, and their semantic enrichment.

The effort reported in this paper represents an evolution of DIRECT in line with the feedback received over the years and discussions raised by experts during the DESIRE workshop, since our final goal is to deliver a unified infrastructure

[5] `http://www.dei.unipd.it/wdyn/?IDsezione=3314&IDgruppo_pass=121`

and environment for data, knowledge, tools, methodologies and the user community in order to advance the experimental evaluation of complex multimedia and multilingual information systems.

3 Conceptual Modeling of the Evaluation Infrastructure

A conceptual schema provides the means for modeling and representing the reality of interest, lays the foundations for the automatic processing and managing of the identified entities, and it is one of the steps of that abstraction process needed in IR evaluation, as discussed in Section 1. In the context of IR evaluation this is particularly important for reducing the human effort required by evaluation activities and to move experimental evaluation from a handicraft process towards a more "industrial" one. Finally, a conceptual schema provides the basis for managing, making accessible, preserving and enriching experimental data over time. This is especially relevant in the IR field since not only are the experimental data the basis for all the subsequent research and scientific production, but they are also extremely valuable from an economic point of view, as discussed in Section 1.

The conceptual schema of the infrastructure is organised into eight functional areas; Figure 1 provides an intuitive representation of them. The remainder of this section provides a brief description of these areas.

Resource Management area: This area supports the interaction between users/groups and the resources handled by the infrastructure. Resources can be actual data adopted in or produced by evaluation activities, e.g. experimental collections or experiment results, as well as the evaluation activities and tasks carried out within them. With the term *resource* we refer to a generic entity that concerns evaluation activities and with which a user or a group of users can interact.

Metadata area: This area supports the description and the enrichment through metadata of the resources handled by the infrastructure.

Evaluation Activity area: This area identifies the core of the infrastructure; it refers to activities aimed at evaluating applications, systems, and methodologies for multimodal and multimedia information access and retrieval. Entities in this area are not limited to traditional evaluation campaigns, but they also include trial and education activities. *Trial* refers to an evaluation activity that may be actively run by, say, a research group, a person or a corporate body for their own interest. This evaluation activity may be or may not be shared with the community of interest; for instance, a trial activity may be the experiments performed to write a research paper or the activities conducted to evaluate a Web application. The *Education* activities allow us to envision evaluation activities carried out for educational purposes.

Experimental Collection area: This area allows us to set up a traditional IR evaluation environment and to manage the different collections made available by the diverse evaluation forums. A classical IR experimental collection

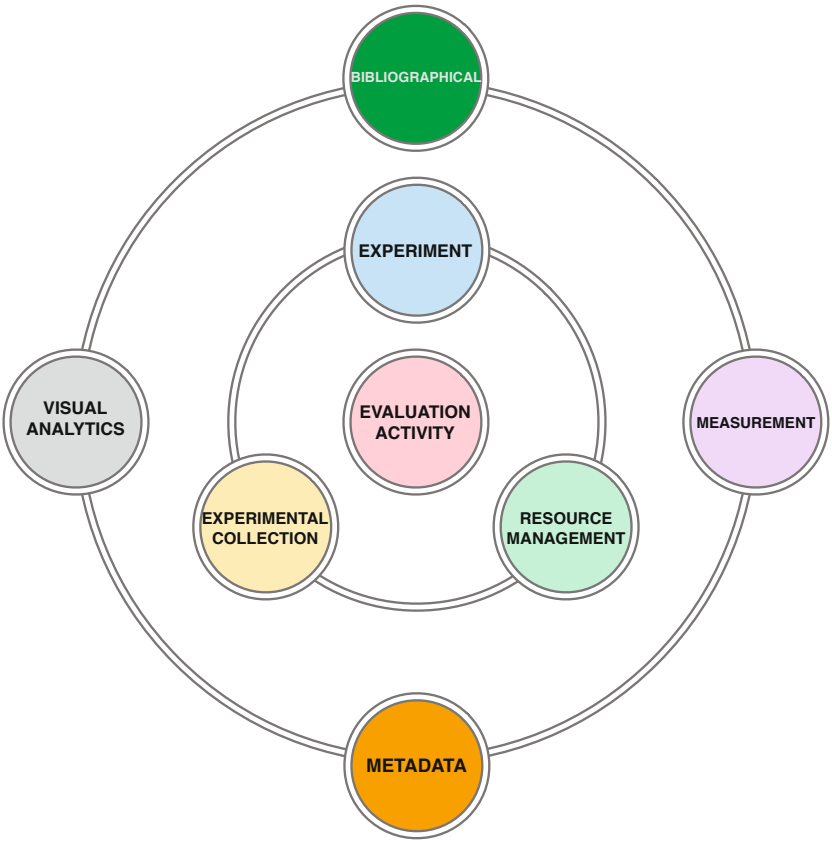


Fig. 1. The conceptual areas of the Evaluation Infrastructure

is a triple composed by a corpus of documents, a group of topics and a set of assessments on the documents with regard to the considered topics. In the abstraction process particular attention has been paid to the the concept of *topic*, because of the diversity of the information needs that have to be addressed in different evaluation tasks.

Experiment area: This area concerns the scientific data produced by an experiment carried out during an evaluation activity. The evaluation infrastructure considers three different types of experiment: run, guerrilla, and living. A *Run* is defined as a ranked list of documents for each topic in the experimental collection [18]. A *Guerrilla* experiment identifies an evaluation activity performed on corporate IR systems (e.g. a custom search engine integrated in a corporate Web site). A *Living* experiment deals with the specific experimental data resulting from the Living Retrieval Laboratories, which will examine the use of operational systems as experimental platform on which to conduct user-based experiments to scale.

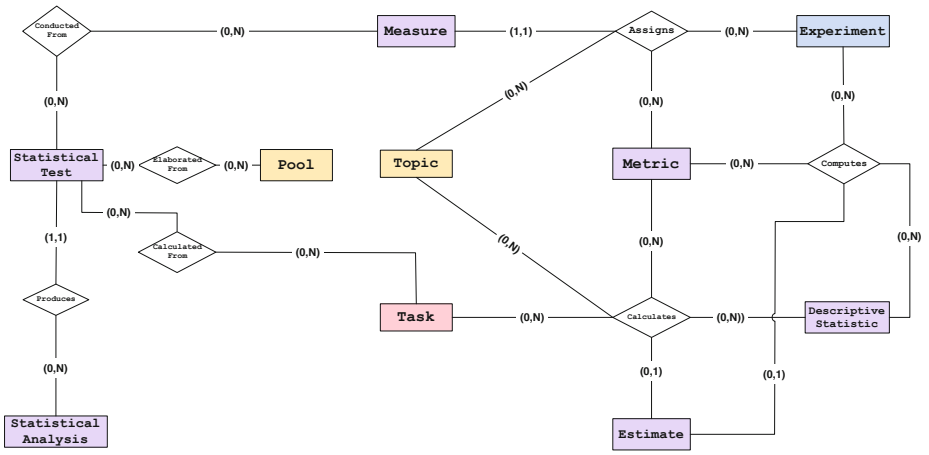


Fig. 2. The ER Schema modeling the Measurement Area

Bibliographical area: This area is responsible for making explicit and retaining the relationship between the data that result from the evaluation activities and the scientific production based on these data.

Measurement area: This area concerns the measures used for evaluation activities.

Visual Analytics area: This area manages the information used by the infrastructure to store and recover whichever visualization of the data that the users do.

In the remainder we will focus on the measurement and visual analytics areas because they provide concrete cases for some of the issues discussed in Section 1.

3.1 Measurement Area

The Measurement area concerns the measures adopted for evaluation activities. Figure 2 shows the ER schema of this area; `Metric` is the main entity and it refers to a standard of measurement allowing us to quantify the effectiveness and the efficiency of a system under evaluation and also to optimize systems themselves. The `Measure` entity represents the value of a `Metric` calculated on some experiments handled by the infrastructure. Other entities in this area are: `Statistical Analysis` which represents a list of the statistical analyses supported by the infrastructure, `Descriptive statistics` which are used to describe the basic features of the data in a study, and `Statistical test` which provides a mechanism for making quantitative decisions about a process or processes. The estimated numerical value of a `Descriptive Statistic` calculated by the infrastructure is represented by the `Estimate` entity.

Figure 2 depicts the relationships among these entities and other entities in the evaluation activity, the experimental collection, and the experiment area,

i.e. Topic, Task and Experiment. For a topic-experiment pair a specific value of a metric, namely a measure, is assigned – i.e. a Measure refers to one and only one Experiment-Topic-Metric triple through the relationship Assigns; an example is the value computed for the metric average precision on the data of an experiment for a specific topic. If we consider the results on an experiment basis, then Descriptive Statistics can be computed for a given Metric – e.g. the Mean Average Precision over all the topics adopted for the Experiment under consideration; this is modeled through the Computes relationship in Figure 2. Descriptive Statistics can be computed also on a task basis, e.g. the variance for a given Topic over all the Experiments submitted for a specific Task; this is modeled by the relationship Calculates that involves the Task, the Metric, the Descriptive Statistic and the Estimate entities.

A Statistical Analysis can produce a value for a specific statistical test; the Statistical Test value can be Elaborated From data in no, one or more Pools, or Calculated From data from no, one or more Tasks, or Computed From an Experiment. Lastly, a Statistical Test value can be obtained by the test Conducted on no, one or more Measures.

The main point here is that explicitly considering the entities in the measurement area as a part of the conceptual schema we are able to retain and make accessible not only experimental data, but also evaluation methodologies and the context wherein metrics and methodologies have been applied. It is our opinion that this is crucial for the definition and the adoption of shared evaluation protocols, which is the main aim of international evaluation initiatives.

3.2 Visual Analytics Area

The Visual Analytics area manages the information used by the infrastructure to store and retrieve parametric and interactive visualization of the data. Indeed, visualizations are not static objects but dynamic ones which are built up via subsequent interactions of the users with the experimental data and the infrastructure. The main entities are: Visualization which refers to the information used by the infrastructure to store a visualization of the data as well as the history of all the interactions of a user with the experimental data, and Snapshot which stores the snapshots of a given visualization. The relationships among the entities in this area are depicted in Figure 3.

Figure 3 also depicts the relationship between the Visualization entity and entities in the Evaluation Activity, the Experimental Collection, the Experiment and the Measurement area. Every visualization can be related to no, one or more Tasks – see relationship ViTa, to no, one or more Pools – see relationship ViPo, to no, one or more Experiments – see relationship ViEx, to no, one or more Statistical Tests – and see relationship ViSt.

4 Architecture of the Evaluation Infrastructure

The architecture of the evaluation infrastructure is based on the introduced conceptual model and stems from an evolution of the DIRECT [16] system.

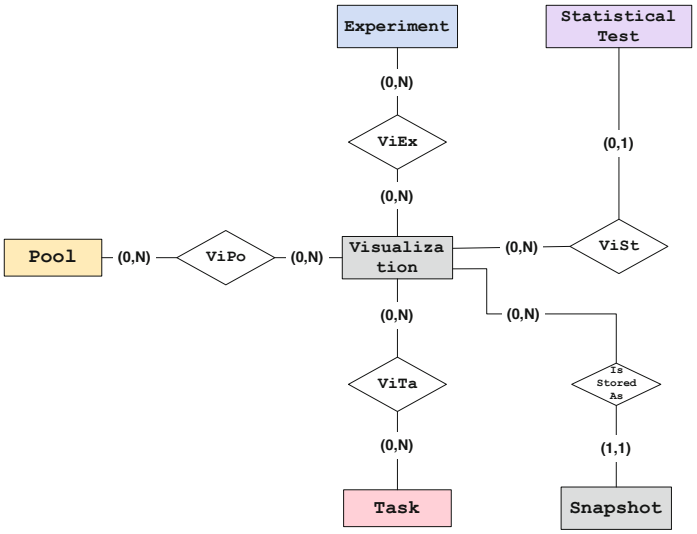


Fig. 3. The relationships between the `Visualization` entity and entities in the Evaluation Activity, the Experimental Collection, the Experiment and the Measurement area

The architecture and the implementation of the system have been developed by exploiting open source technologies, software and frameworks, in order to guarantee a platform which is cooperative, modular, scalable, sustainable over time and allowing interoperability among different systems.

Figure 4 shows the architecture of the system. The right stack summarizes the layers modeling the application, while the left stack shows the building blocks of the implementation of the system. At the lowest levels of the stack – see point (1) of Figure 4 – data stored into database and indexes are mapped to resources and vice versa. The communication with the upper levels is granted through the mechanism of the *Data Access Object (DAO)* pattern. The application logic layer is in charge of the high-level tasks made by the system, such as the enrichment of raw data, the calculation of metrics and the carrying out of statistical analyses on experiments. These resources, shown at point (2), are therefore accessible by remote devices via HTTP through a *REpresentational State Transfer (REST)*ful Web service, represented by points (3) and (5).

The Access Control Infrastructure, point (4), takes care of monitoring the various resources and functionalities offered by the system. It performs authentication by asking for user credentials to log it into the system, and authorization by verifying if the logged in user requesting an operation holds sufficient rights to perform it. The logging infrastructure, which lays behind all the components of the DIRECT system, captures information such as the user name, the *Internet Protocol (IP)* address of the connecting host, the action invoked by the user, the messages exchanged among the components of the system, and any

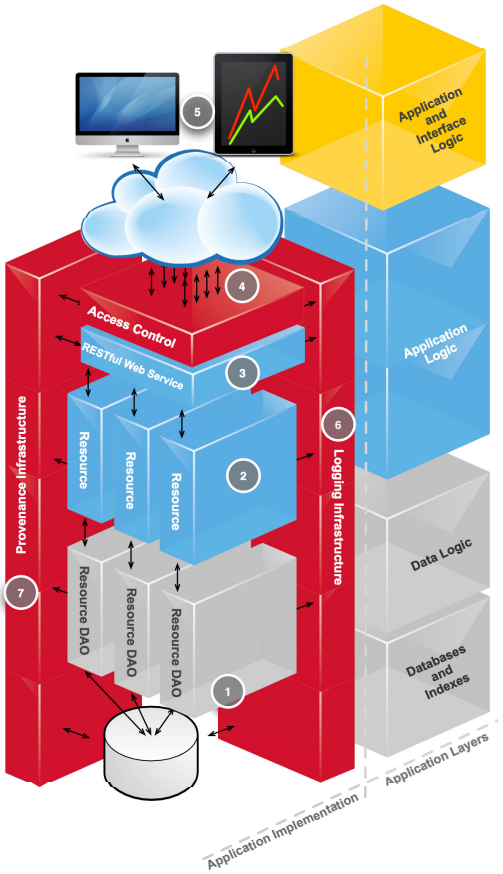


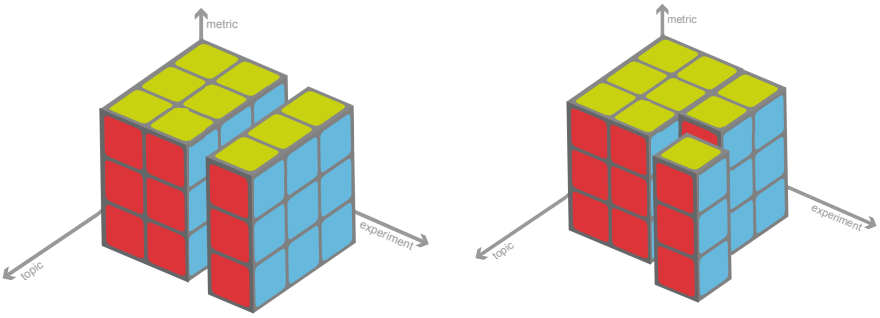
Fig. 4. The Architecture of the DIRECT System as a REST Web Service

error condition, if necessary. The Provenance Infrastructure – point (7) in Figure 4 – is in charge of keeping track of the full lineage of each resource managed by the system since its first creation, allowing granted users to reconstruct its full history and modifications over time.

Next section will focus on the RESTful Web Service level (3) and reports on a use case scenario for accessing experimental data when considering the inter-area involving topics, experiments and metrics.

5 Use Case Scenario: Tasks, Experiments, and Metrics

This use case scenario describes how the users of the DIRECT system can access experimental data about task, experiments, and related metrics in order to process them.



(a) (b)

Fig. 5. Topics, Experiments, and Metrics Data Matrix Sliced on a fixed Experiment (Figure 5a), and for a fixed Topic-Experiment pair (Figure 5b).

An HTTP GET request for a task with identifier `id_tsk` and namespace `ns_tsk`:

`/task/{id_tsk};{ns_tsk}/metric`

will provide data about topics, experiments, and metrics as response, which can be thought of as the three-dimensional matrix, or *Online Analytical Processing (OLAP)* data cubes [19], as those sketched in Figure 5.

The data cube can be rotated (pivot operation) to show topics, experiments and metrics as rows or columns, providing alternative visualizations of data that the user can save and export as snapshots. It is also possible to select and reorder rows or columns, and slice portions of cube, as shown in Figure 5.

Let us consider the case of a user that is interested in an experiment, specifically how an approach performs on each single topic when considering all the distinct metrics made available by the infrastructure. Figure 5a shows the slice of the OLAP Data Cube that can interest this user. The HTTP GET request provided to the DIRECT system to gather all the data about an experiment will be: the URL `http://direct.dei.unipd.it/` followed by

`task/{id_tsk};{ns_tsk}/experiment/{id_exp};{ns_exp}/metric`

For each slice it is possible to refine the request specifying two parameters instead of one, then obtaining a single column from the sliced data cube. For example, if the user is interested in the system performance on a specific topic, the HTTP GET request will be:

`task/{id_tsk};{ns_tsk}/topic/{id_tpt};{ns_tpc}`
`/experiment/{id_exp};{ns_exp}/metric`

The response of this request corresponds to a single column of the data cube that provides information for all the metrics for a given topic in the context of a given experiment – see Figure 5b. Another example could involve a track

coordinator who is writing a track overview paper. He could be interested in the performance of diverse systems, e.g. those that participated to a specific task in a track, for a specific topic; in that case the slice of interest is the one that can be obtained for a fixed topic by the following request

`task/{id_tsk};{ns_tsk}/topic/{id_tpt};{ns_tpt}/metric`

6 Final Remarks

In this paper we discussed the motivations and presented the specification of an IR evaluation infrastructure. We described its underlying conceptual schema, its architecture, and the API to interact with the resources it manages.

Besides supporting the design of an innovative evaluation infrastructure, another goal of this work is to propose a common abstraction of IR evaluation activities that can be exploited to: (i) share and re-use the valuable scientific data produced by experiments and analysis, (ii) employ innovative and interactive visual analytics techniques in IR experimental evaluation, and (iii) envision evaluation activities other than traditional IR campaigns.

Acknowledgements. The authors wish to thanks Marco Dussin for his contributions to the definition of the infrastructure to which he has contributed while he was part of the team of the University of Padua. The authors wish to thanks all PROMISE partners for the useful discussions on many aspects related to the evaluation infrastructure.

The work reported in this paper has been supported by the PROMISE network of excellence (contract n. 258191) project as a part of the 7th Framework Program of the European commission (FP7/2007-2013).

References

1. Harman, D.K.: Information Retrieval Evaluation. Morgan & Claypool Publishers, USA (2011)
2. Rowe, B.R., Wood, D.W., Link, A.L., Simoni, D.A.: Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program. RTI Project Number 0211875, RTI International, USA (2010), `http://trec.nist.gov/pubs/2010.economic.impact.pdf`
3. Allan, J., et al.: Frontiers, Challenges, and Opportunities for Information Retrieval – Report from SWIRL 2012. In: The Second Strategic Workshop on Information Retrieval in Lorne, SIGIR Forum, vol. 46 (in print, February 2012)
4. Sanderson, M.: Test Collection Based Evaluation of Information Retrieval Systems. Foundations and Trends in Information Retrieval (FnTIR) 4, 247–375 (2010)
5. Armstrong, T.G., Moffat, A., Webber, W., Zobel, J.: Improvements That Don't Add Up: Ad-Hoc Retrieval Results Since 1998. In: Proc. 18th International Conference on Information and Knowledge Management (CIKM 2009), pp. 601–610. ACM Press, New York (2009)
6. Zhang, J.: Visualization for Information Retrieval. Springer, Heidelberg (2008)

7. Newman, D., Baldwin, T., Cavedon, L., Huang, E., Karimi, S., Martínez, D., Scholer, F., Zobel, J.: Visualizing Search Results and Document Collections Using Topic Maps. Journal of Web Semantics 8, 169–175 (2010)
8. Banks, D., Over, P., Zhang, N.F.: Blind Men and Elephants: Six Approaches to TREC data. Information Retrieval 1, 7–34 (1999)
9. Sormunen, E., Hokkanen, S., Kangaslampi, P., Pyy, P., Sepponen, B.: Query performance analyser: a web-based tool for ir research and instruction. In: Järvelin, K., Beaulieu, M., Baeza-Yates, R., Hyon Myaeng, S. (eds.) Proceedings of SIGIR 2002, p. 450. ACM, New York (2002)
10. Keim, D.A., Mansmann, F., Schneidewind, J., Ziegler, H.: Challenges in Visual Data Analysis. In: Banissi, E. (ed.) Proc. of the 10th International Conference on Information Visualization (IV 2006), pp. 9–16. IEEE Computer Society, Los Alamitos (2006)
11. Di Buccio, E., Dussin, M., Ferro, N., Masiero, I., Santucci, G., Tino, G.: To Rerank or to Re-query: Can Visual Analytics Solve This Dilemma? In: Forner, P., Gonzalo, J., Kekäläinen, J., Lalmas, M., de Rijke, M. (eds.) CLEF 2011. LNCS, vol. 6941, pp. 119–130. Springer, Heidelberg (2011)
12. Ferro, N., Sabetta, A., Santucci, G., Tino, G.: Visual Comparison of Ranked Result Cumulated Gains. In: Miksch, S., Santucci, G. (eds.) Proc. 2nd International Workshop on Visual Analytics (EuroVA 2011), pp. 21–24. Eurographics Association, Goslar (2011)
13. European Union: Riding the wave. How Europe can gain from the rising tide of scientific data. Printed by Osmotica.it, Final report of the High level Expert Group on Scientific Data (2010)
14. Agosti, M., Di Nunzio, G.M., Ferro, N.: Scientific Data of an Evaluation Campaign: Do We Properly Deal with Them? In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 11–20. Springer, Heidelberg (2007)
15. 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)
16. Dussin, M., Ferro, N.: Managing the Knowledge Creation Process of Large-Scale Evaluation Campaigns. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 63–74. Springer, Heidelberg (2009)
17. Agosti, M., Ferro, N., Thanos, C.: DESIRE 2011: First international workshop on data infrastructures for supporting information retrieval evaluation. In: Proc. of the 20th ACM International Conference on Information and Knowledge Management, pp. 2631–2632. ACM, New York (2011)
18. Voorhees, E.M., Harman, D.K.: TREC: Experiment and Evaluation in Information Retrieval. The MIT Press, MA (2005)
19. Elmasri, R., Navathe, S.B.: Fundamentals of Database Systems, 4th edn. Addison Wesley, Reading (2003)