An Innovative Approach to Data Management and Curation of Experimental Data Generated through IR Test Collections

Maristella Agosti, Giorgio Maria Di Nunzio, Nicola Ferro and Gianmaria Silvello

Abstract This paper describes the steps that led to the invention, design and development of the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* system for managing and accessing the data used and produced within experimental evaluation in *Information Retrieval (IR)*. We present the context in which DIRECT was conceived, its conceptual model and its extension to make the data available on the Web as *Linked Open Data (LOD)* by enabling and enhancing their enrichment, discoverability and re-use. Finally, we discuss possible further evolutions of the system.

1 Introduction

Experimental evaluation is a fundamental topic of *Information Retrieval (IR)* and it has the Cranfield paradigm (Cleverdon, 1997) at its core. The two key components of experimental evaluation are experimental collections and evaluation campaigns organized at an international level. The management of experimental collections – i.e. documents, topics and relevance judgments – and of the data produced by the evaluation campaigns – i.e. runs, measures, descriptive statistics, papers and reports – are of central importance to guarantee the possibility of conducting evaluation experiments that are repeatable and that permit re-usability of the collections.

A crucial aspect for IR evaluation is to ensure the best exploitation and interpretation, over large time spans, of the used and produced experimental data. Nev-

Maristella Agosti; Department of Information Engineering, University of Padua, Italy; e-mail: maristella.agosti@unipd.it
Giorgio Maria Di Nunzio; Department of Information Engineering, University of Padua, Italy; e-mail: giorgiomaria.dinunzio@unipd.it
Nicola Ferro; Department of Information Engineering, University of Padua, Italy; e-mail: nicola.ferro@unipd.it
Gianmaria Silvello; Department of Information Engineering, University of Padua, Italy; e-mail: gianmaria.silvello@unipd.it

ertheless, this aspect has often been overlooked in the field, since researchers are generally more interested in developing new algorithms and methods rather than modeling and managing the experimental data (Agosti et al, 2007b,c).

As a consequence, within the *Conference and Labs of the Evaluation Forum (CLEF)* evaluation campaigns, we worked on modeling the IR experimental data and on designing a research infrastructure able to manage, curate and grant access to them. This effort led to the invention, design and development of the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* system (Agosti et al, 2012; Ferro et al, 2011) and it raised awareness of the importance of curating and managing research data in the community and beyond (Agosti et al, 2009; Allan et al, 2012; Zobel et al, 2011; Agosti et al, 2013, 2014).

DIRECT enables the typical IR evaluation workflow, and manages the scientific data used and produced during large-scale evaluation campaigns. In addition, DIRECT has the potential to support the archiving, access, citation, dissemination and sharing of the experimental results.

On the top of DIRECT, we successively added some *Linked Open Data (LOD)* functionalities (Heath and Bizer, 2011) – i.e. the LOD-DIRECT system – to enable the discoverability, enrichment and the interpretability of the experimental data. We defined a *Resource Description Framework (RDF)* model of the IR scientific data also modelling their connections with the scientific papers related and based on them. We also provided a methodology for automatically enriching the data by exploiting relevant external entities from the LOD cloud (Silvello et al, 2017).

The paper is organized as follows: Section 2 introduces the complex and rich field of experimental evaluation and the Cranfield paradigm, and provides the scientific context in which DIRECT has been invented, designed and developed. Section 3 presents the conceptual model of the infrastructure, and the main conceptual areas composing it, highlighting how experimental data are modeled within the system. Section 4 describes the semantic model defined for publishing IR experimental data on the Web as LOD and LOD-DIRECT. Section 5 refers to related work. Finally, Section 6 discusses and considers possible future developments.

2 The Cranfield Paradigm and the Evaluation Campaigns

2.1 Abstraction of IR Systems Evaluation

The evaluation of information retrieval systems is an abstraction of the retrieval process based on a set of choices that represent certain aspects of the real world (directly or indirectly) and ignore others (Robertson, 2008). This abstraction allows researchers in IR to control some of the variables that affect retrieval performance and exclude other variables that may affect the noise of laboratory evaluation (Voorhees, 2002). The “Cranfield paradigm” is at the heart of the design of laboratory experiments of evaluation of information retrieval tools and systems (Cleverdon, 1997;

Harman, 2011). This paradigm defines the notion of the methodology of experimentation in IR, where the goal is to create "a laboratory type situation where, freed as far as possible from the contamination of operational variables, the performance of index languages could be considered in isolation" (Cleverdon, 1997). The core of this methodology abstracts away from the details of particular tasks and users and instead focuses on a benchmark called "test collection" which consists of three components: a set of documents, a set of topics, and a ground truth, i.e. a set of relevance assessments for each document-topic pair. The abstracted retrieval task is to rank the document set for each topic, then the effectiveness of a system for a single topic is computed as a function of the ranks of the relevant documents (Voorhees, 2007).

Some years after the Cranfield paradigm was established, researchers in the field of IR noted that the collections existing at that time, which had been designed and created for a specific experimental evaluation of a system and/or a comparison between systems, were re-used for many other experiments, for which they were not ideal (Spärck Jones and van Rijsbergen, 1975; Spärck Jones and Bates, 1977). Some of the issues were related to the lack of suitable test data and the way that the experiments were documented often without suitable caveats. Quoting a passage from (Spärck Jones and van Rijsbergen, 1975):

> "There is a widespread feeling among research workers that existing test collections are inadequate because they are small and/or careless and/or inappropriate. They may also not be fully machine-readable, or may be in an esoteric machine format."

On the basis of these considerations, Karen Spärck Jones and Keith van Rijsbergen clarified and illustrated the characteristics that an 'ideal' test collection must have to overcome the aforementioned problems (Spärck Jones and van Rijsbergen, 1975; Robertson, 2008).

2.2 The Ideal Test Collection and TREC

The concept of an ideal test collection was implemented for the first time in the context of the first *Text REtrieval Conference (TREC)*[1] in 1992, that is many years after its definition. One of the goals of TREC has been to provide a shared task evaluation that allows cross-system comparisons. In addition to the initial traditional ad-hoc task, there have been a wide range of experimental tracks that have focused on new areas or particular aspects of text retrieval since TREC 4 (Harman, 1995). To adhere to the ideal test collection characteristics, for each TREC track participants receive: a collection of documents obtained from some external source; a collection of topics, which may also be obtained externally or may be created internally; and a set of relevance assessments, known as *qrels*. Each participant tests a search system on the collection and produces as a result a ranked list of documents for each topic –

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

known as a *run* – which is submitted to NIST. The runs submitted by the participants are pooled in order to produce the set of relevance assessment (Robertson, 2008).

Since its beginning in 1992, the TREC effort has had a profound influence on all aspects of evaluation, from the formatting of test collection documents, topics, and qrels, through the types of information needs and relevance judgments made, to the precise definition of evaluation measures used (Voorhees and Harman, 2005; Sanderson, 2010). In addition to experimental collection material, TREC has also greatly encouraged the development of good methods of experimentation. The standard of rigour of experimental methodology has been vastly improved (Robertson, 2008) thanks to TREC, which has become a yearly evaluation initiative, or evaluation campaign, of reference for all academic and industrial communities of information retrieval.

2.3 The Management of Data Produced in the Context of Evaluation Campaigns

Donna Harman and her colleagues appeared to be the first to realize that if the documents and topics of a test collection were distributed for little or no cost, a large number of groups would be willing to use that data in their search systems and submit runs back to TREC at no cost (Sanderson, 2010). Moreover, the materials and methods TREC has generated are materials and methods for laboratory experiments (Robertson, 2008). In this respect, since its beginning TREC has promoted the concept of reusability which facilitates research.

Despite the fact that IR has traditionally been very rigorous about experimental evaluation, researchers in this field have raised some concerns about the reproducibility of system experiments because, among other things, there is not a clear methodology for managing experimental data across different conferences and evaluation initiatives (Ferro, 2017). In fact, after TREC, other evaluation campaigns were launched to deal with the evaluation of many different IR approaches and systems that were being defined also thanks to the development of many different types of IR systems and tools, such as, for example, Web search engines.

Some important relevant evaluation campaigns that have been launched over the years and that are still active now are: NTCIR (NII Testbed and Community for Information access Research), Japan, from 1999[2]; CLEF (Conference and Labs of the Evaluation Forum), Europe, from 2000[3]; FIRE (Forum for Information Retrieval Evaluation), India, from 2008[4].

The *INitiative for the Evaluation of XML Retrieval (INEX)* has provided the means to evaluate focused retrieval search engines, especially *eXtensible Markup*

[2] `http://research.nii.ac.jp/ntcir/index-en.html`

[3] `http://www.clef-initiative.eu/`

[4] `http://fire.irsi.res.in/`

Language (XML) retrieval; it was launched in 2002, came under the CLEF umbrella in 2012, but ran for the last time in 2014[5].

As reported, many evaluation initiatives are active and produce important results for the evaluation of IR systems and tools. Naturally, these different initiatives have been launched and conducted to respond to different research questions, so they have specificities that do not always make cross-comparability between the initiatives possible. As a consequence, the produced experimental results are often not cross comparable. We started from this consideration to work on proposing a conceptual model of an infrastructure that would face and solve some of the problems related to the management and curation of the data produced during an evaluation campaign (Agosti et al, 2007c,a).

3 Conceptual Model of the Infrastructure

In IR, as well as in other related scientific fields, a crucial topic that has to be addressed is how to guarantee that the data produced by the scientific activities are consistently managed, are made accessible and available for re-use and are documented to make them easily interpretable. In IR evaluations, these are key aspects, and especially in the context of large evaluation campaigns such as CLEF. For example, the importance of describing and annotating scientific datasets is discussed in (Bowers, 2012), noting that this is an essential step for the interpretation, sharing, and reuse of the datasets.

We thus began an exercise aimed at modeling the IR experimental data and designing a software infrastructure able to manage and curate them, which led to the development of the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* system (Di Nunzio and Ferro, 2005; Agosti et al, 2012). This effort contributed to raising awareness and consensus in the research community and beyond (Agosti et al, 2009; Allan et al, 2012; Forner et al, 2013; Zobel et al, 2011; Ferro et al, 2011).

DIRECT models all the aspects of a typical evaluation workflow in IR and provides the means to deal with some advanced aspects that have been receiving attention in recent years, such as bibliometrics based on data and the visualization of scientific data.

We can model the main phases of the IR experimental evaluation workflow as follows:

- The first phase regards the creation of the experimental collection composed of the acquisition and preparation of the documents (D) and the creation of topics (T) from which a set of queries is generated.
- The second phase concerns the participants in the evaluation campaign who run experiments and test their systems.

[5] `https://inex.mmci.uni-saarland.de/`

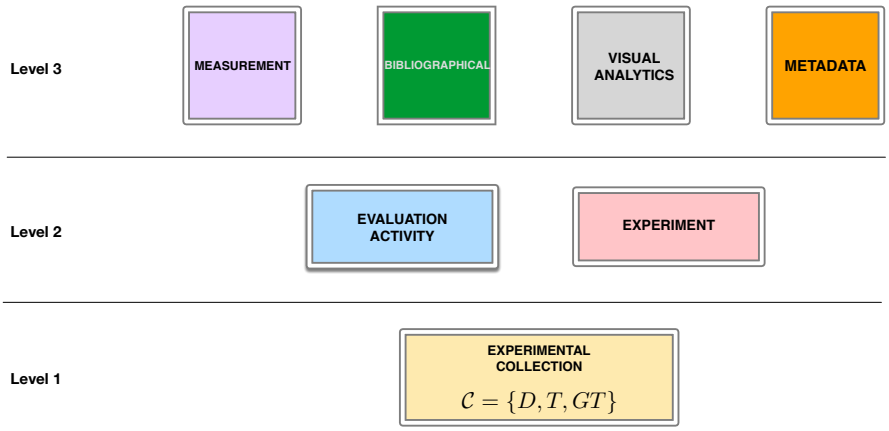

Fig. 1 The conceptual areas of the evaluation infrastructure.

- In the third phase, the experiments are gathered and used by the campaign organizers to create the ground-truth (GT).
- In the fourth phase, measurements are calculated.
- In the fifth phase the measurements are used to produce descriptive statistics and conduct statistical tests about the behavior of one or more systems.
- The sixth and last phase regards the scientific production where both participants and organizers prepare reports about the campaign and the experiments, the techniques they used, and their findings. This phase usually continues also after the conclusion of the campaign as the investigations of the experimental results require a deeper understanding and further analyses which may lead to the production of conference and journal papers.

The conceptual schema of the infrastructure, abstracting from the actual phases of the IR experimental evaluation workflow, models the evaluation workflow by means of seven functional areas organized in three main conceptual levels; Figure 1 provides an intuitive representation of them. The three levels are built one on top of the other since the experimental collection area constitutes the basis of the evaluation activities and the experiments on level 2. In the same fashion, the measurement, bibliographical, visual analytics and metadata areas on level 3 depend on the areas on level 2.

We document in the following the aim and the content of each functional area.

Experimental Collection area: This area belongs to the first conceptual level and it allows us to set up a traditional IR evaluation environment following the classic Cranfield paradigm based on the triple $\mathscr{C} = \{D, T, GT\}$: 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.

Evaluation Activity area: This area belongs to the second conceptual level and builds on the experimental collection area. It 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 go beyond the traditional evaluation campaigns by including 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 or may not be shared with the community of interest; for instance, a trial activity may be the experiments performed to answer a research question and 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. In a certain sense, this area extends the activities considered by the Cranfield paradigm.

Experiment area: This area belongs to the second conceptual level and concerns the scientific data produced by an experiment carried out during an evaluation activity. Also in this case, this area models the traditional Cranfield experimental settings and extends it by allowing other side evaluation activities. Indeed, the evaluation infrastructure considers three different types of experiment: run, guerrilla, and living. A *Run*, produced by an IR system, is defined as a ranked list of documents for each topic in the experimental collection (Voorhees and Harman, 2005) in a classic IR evaluation context. A *Guerrilla* experiment identifies an evaluation activity performed on corporate IR systems (e.g. a custom search engine integrated in a corporate Web site) (Agosti et al, 2012); in a guerrilla experiment, the evaluation process is defined by a set of experimental activities aimed at assessing different aspects of the application, such as the completeness of the index of an ad-hoc search engine or the effectiveness of the multilingual support. For this reason the evaluation metrics may differ from those used during a Run experiment. A *Living* experiment deals with the specific experimental data resulting from the Living Retrieval Laboratories, which examines the use of operational systems on an experimental platform on which to conduct user-based experiments to scale.

Measurement area: This area belongs to the third conceptual level and concerns the measures used for evaluation activities. This area is one of the most important of the infrastructure and it constitutes one element of distinction between DIRECT and other modeling efforts in the IR evaluation panorama. In Figure 2 we can see relationships among the main entities of this area and other entities in the evaluation activity, the experimental collection, and the experiment area. 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`. If we consider the results on an experiment basis, then `Descriptive Statistics` can be computed for a given `Metric`. `Descriptive Statistics` can be computed also on a task basis. A `Statistical Analysis` can produce a value for a specific statistical test; the `Statistical Test` value can be `Elaborated From`

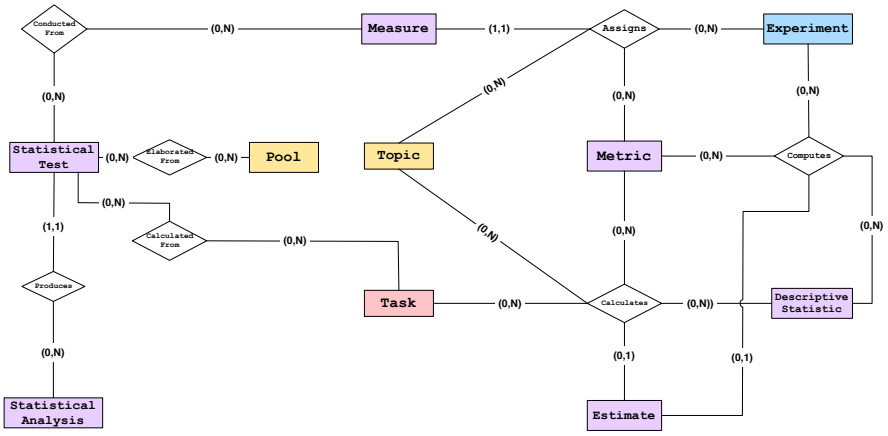

Fig. 2 The ER schema modeling the measurement area.

data in none, one or more `Pools`, or `Calculated From` data from none, one or more `Tasks`, or `Computed From` an `Experiment`.

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.

Metadata area: This area belongs to the third level and supports the description and the enrichment through metadata of the resources handled by the infrastructure. Generally, metadata describing the resources are not considered central resources in an evaluation infrastructure: whereas, in DIRECT they are considered important resources and managed alongside the other classical evaluation resources. This allows us to use metadata in concert with measures and experiments for enriching the experimental data as we discuss below.

Bibliographical area: This area belongs to the third level and it 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. This area is central for dealing with bibliometrics of experimental data and for dealing with data provenance (Buneman et al, 2000) and citation (Davidson et al, 2017),

Visual Analytics area: This area belongs to the third level and it manages the information used by the infrastructure to store and recover whatever visualization of the data that users produce. This area manages the information used by the infrastructure to store and retrieve parametric and interactive visualizations of the data.

To the best of our knowledge, DIRECT is the most comprehensive tool for managing all the aspects of the IR evaluation methodology, the experimental data produced and the connected scientific contributions. Besides supporting the design of an innovative evaluation infrastructure, another goal of DIRECT is to provide a com-

mon abstraction of IR evaluation activities that can be exploited to share and re-use the valuable scientific data produced by experiments and analysis and to envision evaluation activities other than traditional IR campaigns.

4 A Semantic Mapping of the Conceptual Model

Research data are of key importance across all scientific fields as these data constitute a fundamental building block of the system of science. Recently, a great deal of attention has been dedicated to the nature of research data and how to describe, share, cite and re-use them in order to enable reproducibility in science and to ease the creation of advanced services based on them (Borgman, 2015; Silvello, 2017). In this context, the *Linked Open Data (LOD)* paradigm (Heath and Bizer, 2011) is a de-facto standard for publishing and enriching data; it allows the opening-up of public data in machine-readable formats ready for consumption, re-use and enrichment through semantic connections enabling new knowledge creation and discovery possibilities. The LOD paradigm can be mainly seen as a method of publishing structured data so that data can be interlinked. It builds upon standard Web technologies such as *HyperText Transfer Protocol (HTTP)* and RDF[6], but rather than using them to serve web pages for humans, "it extends them to share information in a way that can be read automatically by machines"[7].

In the field of IR, the LOD paradigm is not as central as it is in other fields such as life science research (Gray et al, 2014) and social sciences (Zapilko et al, 2013). So, despite the centrality of data, in IR there are no shared and clear ways to publish, enrich and re-use experimental data as LOD with the research community.

To target this aspect of data sharing, re-use and enrichment within the DIRECT infrastructure, we defined an RDF model (W3C, 2004) for representing experimental data and publishing them as LOD on the Web. This can enable seamless integration of datasets produced by different experimental evaluation initiatives as well as the standardization of terms and concepts used to label data across research groups and interested organizations (Silvello et al, 2017).

Moreover, with the purpose of augmenting the access points to the data as well as the potential for their interpretability and re-usability, we built upon the proposed RDF model to automatically find topics in the scientific literature, exploiting the scientific IR data as well as connecting the dataset with other datasets in the LOD cloud.

The detection of scientific topics related to the data produced by the experimental evaluation and the enrichment of scientific data mainly concerns the "experiment area" and areas of the scientific production (level 3) of the evaluation infrastructure. Regarding the experimental evaluation and the scientific production area, the con-

[6] `https://www.w3.org/standards/semanticweb/data`

[7] `https://en.wikipedia.org/wiki/Linked_data`

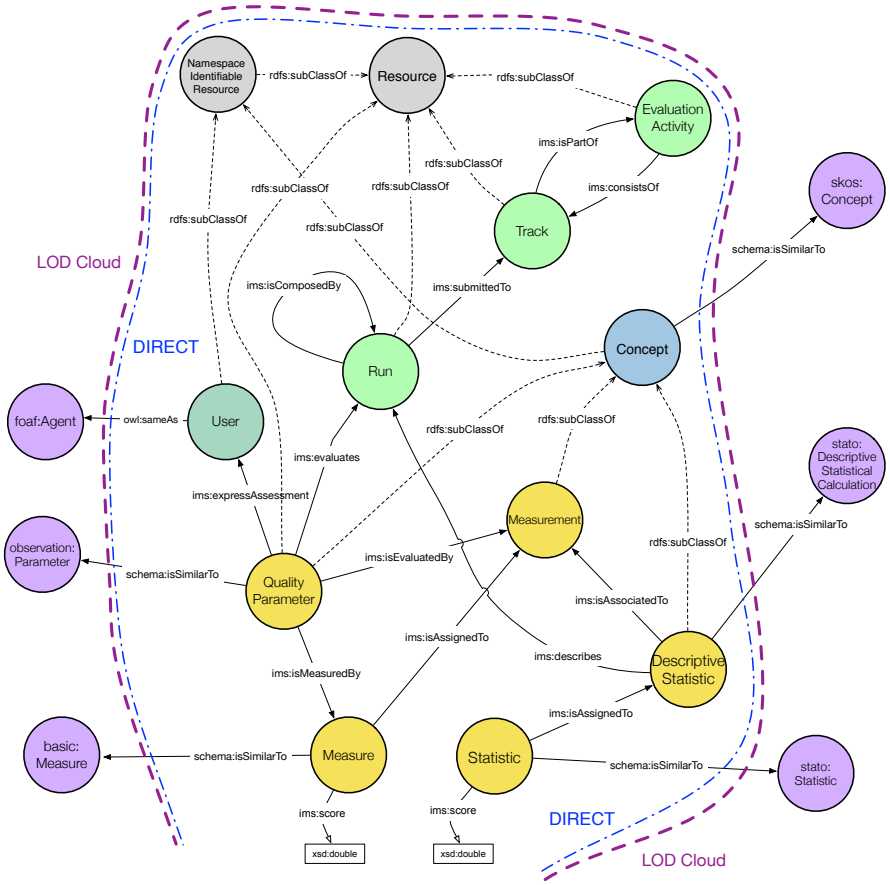

Fig. 3 The experiment area classes and properties.

ceptual model of DIRECT has been mapped into an RDF model and adopted for enriching and sharing the data produced by the evaluation activities.

In Figure 3 we can see the classes and properties of the experiment area as reported and described in (Silvello et al, 2017). Please note that the IMS namespace in this case indicates that all the class and property names are defined within the DIRECT workspace; this enables the distinction with other classes and properties in the LOD cloud which may have the same denomination, but of course different namespace. The area shown in the figure is central to the DIRECT infrastructure and it is connected to the most important resources for the evaluation activities. Hence, we focus on this to present the semantic model we designed, even though it encompasses almost all the areas described above.

The experiment area can be divided into two main parts: one comprising the `Run`, `Track` and `Evaluation Activity` classes modeling the experiments and the other one comprising the `Quality Parameter`, `Measurement`, `Measure`,

`Descriptive Statistic` and `Statistic` classes modeling the evaluation of the experiments.

The first part allows us to model an evaluation campaign composed of several runs submitted to a track which is part of an evaluation activity. The second part allows us to model the measurements and the descriptive statistics calculated from the runs and it is built following the model of quality for *Digital Library (DL)* defined by the DELOS Reference Model (Candela et al, 2007) which is a high-level conceptual framework that aims at capturing significant entities and their relationships within the digital library universe with the goal of developing more robust models of it; we extended the DELOS quality model and we mapped it into an RDF model. A `Quality Parameter` is a `Resource` that indicates, or is linked to, performance or fulfilment of requirements by another Resource. A `Quality Parameter` is evaluated by a `Measurement`, is measured by a `Measure` assigned according to the `Measurement`, and expresses the assessment of a `User`. With respect to the definition provided by the *International Organization for Standardization (ISO)*, we can note that: the "set of inherent characteristics" corresponds to the pair (`Resource`, `Quality Parameter`); the "degree of fulfillment" fits in with the pair (`Measurement`, `Measure`); finally, the "requirements" are taken into consideration by the assessment expressed by a `User`.

`Quality Parameter`s allow us to express the different facets of evaluation. In this model, each `Quality Parameter` is itself a `Resource` and inherits all its characteristics, such as, for example, the property of having a unique identifier. `Quality Parameter`s provide information about how, and how well, a resource performs with respect to some viewpoint. They express the assessment of a `User` about the `Resource` under examination. They can be evaluated according to different `Measurement`s, which provide alternative procedures for assessing different aspects of a `Quality Parameter` and assigning it a value, i.e. a `Measure`. Finally, a `Quality Parameter` can be enriched with metadata and annotations. In particular, the former can provide useful information about the provenance of a `Quality Parameter`, while the latter can offer the possibility to add comments about a `Quality Parameter`, interpreting the obtained values, and proposing actions to improve it.

One of the main `Quality Parameters` in relation to an information retrieval system is its effectiveness, meant as its capability to answer user information needs with relevant items. This `Quality Parameter` can be evaluated according to many different `Measurement`s, such as precision and recall (Salton and McGill, 1983). The actual values for precision and recall are `Measure`s and are usually computed using standard tools, such as `trec_eval`[8], which are `Users`, but in this case not human ones.

The `Descriptive Statistic` class models the possibility of associate statistical analyses to the measurements; for instance, a classical descriptive statistic in IR is *Mean Average Precision (MAP)* which is the mean over all the topics of a run of the *Average Precision (AP)* measurement which is calculated topic by topic.

[8] `http://trec.nist.gov/trec_eval/`

The described RDF model has been realized and implemented in the DIRECT system. This allows for accessing the experimental evaluation data enriched by the expert profiles that are created by means of the techniques that will be described in the next sections. This system is called LOD-DIRECT and it is accessible at the URL: `http://lod-direct.dei.unipd.it/`

The data currently available include the contributions produced by the CLEF evaluation activities, the authors of the contributions, information about CLEF tracks and tasks, provenance events and the above described measures. Furthermore, this data has been enriched with expert profiles and topics which are available as linked data as well.

LOD-DIRECT serializes and allows access to the defined resources in several different formats such as XML, JSON, RDF+XML, Turtle[9] and Notation3 (n3)[10].

LOD-DIRECT comes with a fine-grained access control infrastructure which monitors the access to the various resources and functionalities offered by the system. Depending on the operation requested, it performs authentication and authorization.

The access control policies can be dynamically configured and changed over time by defining roles, i.e., groups of users, entitled to perform given operations. This allows institutions to define and put in place their own rules in a flexible way according to their internal organization and working practices. The access control infrastructure allows us to manage the experimental data which cannot be publicly shared such as log files coming from search engine companies.

4.1 Use case

In Figure 4 we can see an example of an RDF graph showing how LOD-DIRECT models topics, author profiles, measures and papers. This use case is taken from (Silvello et al, 2017).

We can see the relationship between a contribution and an author enriched by expertise topics, expert profiles and connections to the LOD cloud. In this figure, we focus on the author (*Jussi Karlgren*) and the contribution (*KarlgrenEtAl-CLEF2012*). Here, there are two main topics, "reputation management" and "information retrieval", which are related to the *KarlgrenEtAl-CLEF2012* contribution. We can see that *KarlgrenEtAl-CLEF2012* is featured by "reputation management" with a score of 0.53 and by "information retrieval" with 0.42, meaning that both these topics are subjects of the contribution; the scores give a measure of how much this contribution is about a specific topic. We can also see that the paper at hand presents the results for the `RepLab 2012` track at `CLEF 2012` where `Gavagai` obtained an `accuracy` of `0.77`.

[9] `http://www.w3.org/TR/turtle/`

[10] `http://www.w3.org/TeamSubmission/n3/`

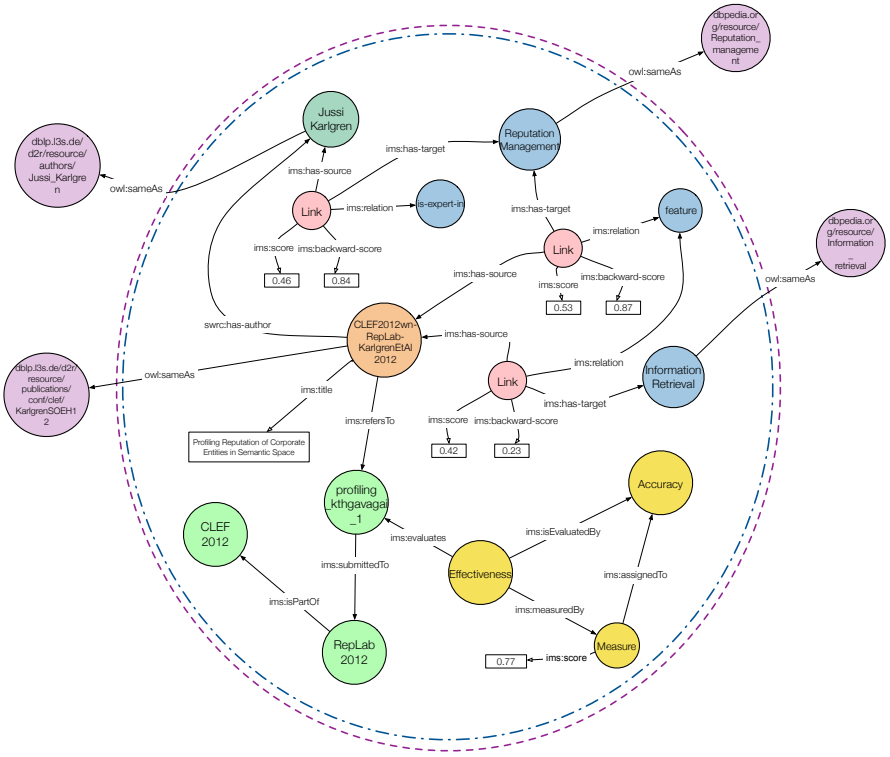

Fig. 4 An example of an RDF graph showing how expertise topics and expert profiles are used for enriching IR experimental data (Silvello et al, 2017).

From this use case we see how LOD-DIRECT models the relationships between papers, authors, topics, measures and evaluation campaigns.

5 Related Work

A crucial question in IR is how to ensure the best exploitation and interpretation of the valuable scientific data employed and produced by the experimental evaluation. To the best of our knowledge, DIRECT is the most comprehensive tool for managing all the aspects of the IR evaluation methodology, the experimental data produced and the connected scientific contributions.

There are other projects with similar goals but with a narrower scope. One is the *Open Relevance Project (ORP)*[11] which is a "small Apache Lucene sub-project aimed at making materials for doing relevance testing for Information Retrieval, Machine Learning and Natural Language Processing into open source"; the goal of

[11] `https://lucene.apache.org/openrelevance/`

this project is to connect specific elements of the evaluation methodology – e.g. experimental collections, relevance judgments and queries – with the Apache Lucene environment in order to ease the work of developers and users. Unfortunately, the project was discontinued in 2014. Moreover, ORP neither considers all the aspects of the evaluation process such as the organization of an evaluation campaign in tracks and tasks or the management of the experiments submitted by the participants to a campaign, nor takes into account the scientific production connected to the experimental data which is vital for the enrichment of the data themselves as well as for the definition of expert profiles.

Another relevant project is EvaluatIR.org[12] (Armstrong et al, 2009) which is focused on the management and comparison of IR experiments. It does not model the whole evaluation workflow and it acts more as a repository of experimental data rather than as an information management system for curating and enriching them.

There are other efforts carried out by the IR community which are connected to DIRECT, even though they have different purposes. One relevant example is the TIRA (TIRA Integrated Research Architecture) Web service (Gollub et al, 2012), which aims at publishing IR experiments as a service; this framework does not take into account the whole evaluation process as DIRECT does and it is more focused on modeling and making available "executable experiments", which is out of the scope of DIRECT. Another relevant system is RETRIEVAL (Ioannakis et al, 2018); this is a web-based performance evaluation platform providing information visualization and integrated information retrieval for the evaluation of IR system. This system has some overlapping features with DIRECT, but it mainly focuses on the evaluation of IR systems rather than on the management of the data produced by evaluation campaigns and the management of the IR evaluation workflow.

6 Discussion

The DIRECT infrastructure effectively supports the management and curation of the data produced during an evaluation campaign. DIRECT has been used since 2005 for managing and providing access to CLEF experimental evaluation data. Over these years, the system has been extended and revised according to the needs and requirements of the community. Currently, DIRECT handles about 35 million documents, more than 13 thousand topics, around 4 million relevance judgments, about 5 thousand experiments and 20 million measures. This data has been used by more than 1,600 researchers from more than 75 countries world-wide.

Thanks to the expertise we have acquired in designing and developing it, we can now say that it would be preferable to have two distinct infrastructures rather than a single one:

- one to manage all those activities which are needed to run a cycle of an evaluation campaign;

[12] `http://wice.csse.unimelb.edu.au:15000/evalweb/ireval/`

- one for the long term preservation and curation of the information produced by the various evaluation campaign cycles over time.

In fact, DIRECT solves two different problems at the same time: those related to the management of the evaluation campaign cycles and those related to the archiving, preservation and curation of the experimental data produced by evaluation campaigns. However, these two kinds of activities are very different and managing them with a single infrastructure adds sizeable complexity to its design and implementation. On the other hand, if two distinct infrastructures were to be designed and implemented, each of them would be focused on a set of more homogeneous activities, resulting in simpler and more effective infrastructures for each specific objective. The results that are collected over time for each individual instance of an evaluation campaign could be used, for example, for activities of data analysis transversal to various periodic evaluation initiatives.

We have considered the possibility of developing two different infrastructures, because we believe that this effort would be extremely useful for the long-term development of the IR area. But developing two distinct infrastructures of this type would involve a significant investment of human and financial resources. Unfortunately, even if there is widespread agreement on the importance of experimental data, this kind of activity is not yet considered mainstream by the IR community. Therefore, to really value the effort and resources needed to implement such infrastructures, the IR community should better acknowledge the scientific value of such endeavours and should conduct them in a coordinated way so as to distribute the effort over different research groups and to produce a coordinated collection of scientific data that is at the same time curated, citable and freely available over the years for future scientific research.

Acknowledgements

The results we have presented have mostly originated in the context of the research activities of the *Information Management System (IMS)* research group of the Department of Information Engineering of the University of Padua, Italy, but they have benefitted from the collaboration and the support of many experts, in particular of Carol Peters of ISTI, CNR, Pisa, Italy, and of Donna Harman of NIST, USA, to whom our sincere thanks are given. The research activities have been supported by the financial support of different European projects, namely DELOS (FP6 NoE, 2004–2007, Contract n. G038-507618), TrebleCLEF (FP7 CA, 2008–2009, Contract n. 215231), and PROMISE (FP7 NoE, 2010–2013, Contract n. 258191).

We are most grateful to our referees for their very helpful comments.

References

Agosti M, Di Nunzio GM, Ferro N (2007a) A Proposal to Extend and Enrich the Scientific Data Curation of Evaluation Campaigns. In: Sakay T, Sanderson M, Evans DK (eds) Proc. 1st International Workshop on Evaluating Information Access (EVIA 2007), National Institute of Informatics, Tokyo, Japan, pp 62–73

Agosti M, Di Nunzio GM, Ferro N (2007b) Scientific Data of an Evaluation Campaign: Do We Properly Deal With Them? In: Peters C, Clough P, Gey FC, Karlgren J, Magnini B, Oard DW, de Rijke M, Stempfhuber M (eds) Evaluation of Multilingual and Multi-modal Information Retrieval : Seventh Workshop of the Cross–Language Evaluation Forum (CLEF 2006). Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 4730, Springer, Heidelberg, Germany, pp 11–20

Agosti M, Di Nunzio GM, Ferro N (2007c) The Importance of Scientific Data Curation for Evaluation Campaigns. In: Thanos C, Borri F, Candela L (eds) Digital Libraries: Research and Development. First International DELOS Conference. Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 4877, Springer, Heidelberg, Germany, pp 157–166

Agosti M, Ferro N, Thanos C (2009) DESIRE 2011: First International Workshop on Data infrastructurEs for Supporting Information Retrieval Evaluation. In: Ounis I, Ruthven I, Berendt B, de Vries AP, Wenfei F (eds) Proc. 20th International Conference on Information and Knowledge Management (CIKM 2011), ACM Press, New York, USA, pp 2631–2632

Agosti M, Di Buccio E, Ferro N, Masiero I, Peruzzo S, Silvello G (2012) DIRECTions: Design and Specification of an IR Evaluation Infrastructure. In: Catarci T, Forner P, Hiemstra D, Peñas A, Santucci G (eds) Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics. Proceedings of the Third International Conference of the CLEF Initiative (CLEF 2012), Lecture Notes in Computer Science (LNCS) 7488, Springer, Heidelberg, Germany, pp 88–99

Agosti M, Fuhr N, Toms E, Vakkari P (2013) Evaluation methodologies in information retrieval (dagstuhl seminar 13441). Dagstuhl Reports 3(10):92–126

Agosti M, Fuhr N, Toms EG, Vakkari P (2014) Evaluation methodologies in information retrieval Dagstuhl seminar 13441. SIGIR Forum 48(1):36–41, DOI 10.1145/2641383.2641390, URL http://doi.acm.org/10.1145/2641383.2641390

Allan J, Aslam J, Azzopardi L, Belkin N, Borlund P, Bruza P, Callan J, Carman C M Clarke, Craswell N, Croft WB, Culpepper JS, Diaz F, Dumais S, Ferro N, Geva S, Gonzalo J, Hawking D, Järvelin K, Jones G, Jones R, Kamps J, Kando N, Kanoulos E, Karlgren J, Kelly D, Lease M, Lin J, Mizzaro S, Moffat A, Murdock V, Oard DW, de Rijke M, Sakai T, Sanderson M, Scholer F, Si L, Thom J, Thomas P, Trotman A, Turpin A, de Vries AP, Webber W, Zhang X, Zhang Y (2012) Frontiers, Challenges, and Opportunities for Information Retrieval – Report from SWIRL 2012, The Second Strategic Workshop on Information Retrieval in Lorne, February 2012. SIGIR Forum 46(1):2–32

Armstrong TG, Moffat A, Webber W, Zobel J (2009) EvaluatIR: an Online Tool for Evaluating and Comparing IR Systems. In: Allan J, Aslam JA, Sanderson M, Zhai C, Zobel J (eds) Proc. 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009), ACM Press, New York, USA, p 833

Borgman CL (2015) Big Data, Little Data, No Data. MIT Press

Bowers S (2012) Scientific workflow, provenance, and data modeling challenges and approaches. Journal on Data Semantics 1(1):19–30, DOI 10.1007/s13740-012-0004-y, URL http://dx.doi.org/10.1007/s13740-012-0004-y

Buneman P, Khanna S, Tan WC (2000) Data provenance: Some basic issues. In: Kapoor S, Prasad S (eds) Foundations of Software Technology and Theoretical Computer Science, 20th Conference, FST TCS 2000 New Delhi, India, December 13-15, 2000, Proceedings., Springer, Lecture Notes in Computer Science, vol 1974, pp 87–93, DOI 10.1007/3-540-44450-5_6, URL https://doi.org/10.1007/3-540-44450-5_6

Candela L, Castelli D, Ferro N, Ioannidis Y, Koutrika G, Meghini C, Pagano P, Ross S, Soergel D, Agosti M, Dobreva M, Katifori V, Schuldt H (2007) The DELOS Digital Library Reference Model. Foundations for Digital Libraries. ISTI-CNR at Gruppo ALI, Pisa, Italy, https://tinyurl.com/y7fxsz2d

Cleverdon CW (1997) The Cranfield Tests on Index Languages Devices. In: Spärck Jones K, Willett P (eds) Readings in Information Retrieval, Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA, pp 47–60

Davidson SB, Buneman P, Deutch D, Milo T, Silvello G (2017) Data citation: A computational challenge. In: Sallinger E, den Bussche JV, Geerts F (eds) Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2017, Chicago, IL, USA, May 14-19, 2017, ACM, pp 1–4, DOI 10.1145/3034786.3056123, URL http://doi.acm.org/10.1145/3034786.3056123

Di Nunzio GM, Ferro N (2005) DIRECT: a Distributed Tool for Information Retrieval Evaluation Campaigns. In: Ioannidis Y, Schek HJ, Weikum G (eds) Proc. 8th DELOS Thematic Workshop on Future Digital Library Management Systems: System Architecture and Information Access, pp 58–63

Ferro N (2017) Reproducibility Challenges in Information Retrieval Evaluation. ACM Journal of Data and Information Quality (JDIQ) 8(2):8:1–8:4, DOI 10.1145/3020206, URL http://doi.acm.org/10.1145/3020206

Ferro N, Hanbury A, Müller H, Santucci G (2011) Harnessing the Scientific Data Produced by the Experimental Evaluation of Search Engines and Information Access Systems. Procedia Computer Science 4:740–749

Forner P, Bentivogli L, Braschler M, Choukri K, Ferro N, Hanbury A, Karlgren J, Müller H (2013) PROMISE Technology Transfer Day: Spreading the Word on Information Access Evaluation at an Industrial Event. SIGIR Forum 47(1):53–58

Gollub T, Stein B, Burrows S, Hoppe D (2012) TIRA: configuring, executing, and disseminating information retrieval experiments. In: Hameurlain A, Tjoa AM, Wagner RR (eds) 23rd International Workshop on Database and Expert Systems Applications, DEXA 2012, Vienna, Austria, September 3-7, 2012, IEEE Computer Society, pp 151–155

Gray AJG, Groth P, Loizou A, Askjaer S, Brenninkmeijer CYA, Burger K, Chichester C, Evelo CTA, Goble CA, Harland L, Pettifer S, Thompson M, Waagmeester A, Williams AJ (2014) Applying Linked Data Approaches to Pharmacology. Architectural Decisions and Implementation. Semantic Web 5(2):101–113

Harman DK (ed) (1995) The Fourth Text REtrieval Conference (TREC-4), National Institute of Standards and Technology (NIST), Special Pubblication 500-236, Washington, USA. http://trec.nist.gov/pubs/trec4/t4_proceedings.html

Harman DK (2011) Information Retrieval Evaluation. Morgan & Claypool Publishers, USA

Heath T, Bizer C (2011) Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool Publishers, USA

Ioannakis G, Koutsoudis A, Pratikakis I, Chamzas C (2018) RETRIEVAL - An Online Performance Evaluation Tool for Information Retrieval Methods. IEEE Trans Multimedia 20(1):119–127, DOI 10.1109/TMM.2017.2716193, URL https://doi.org/10.1109/TMM.2017.2716193

Robertson SE (2008) On the history of evaluation in IR. Journal of Information Science 34(4):439–456, DOI 10.1177/0165551507086989, URL https://doi.org/10.1177/0165551507086989

Salton G, McGill MJ (1983) Introduction to Modern Information Retrieval. McGraw-Hill, New York, USA

Sanderson M (2010) Test Collection Based Evaluation of Information Retrieval Systems. Foundations and Trends in Information Retrieval (FnTIR) 4(4):247–375

Silvello G (2017) Theory and Practice of Data Citation. Journal of the Association for Information Science and Technology (JASIST) p 24 pages

Silvello G, Bordea G, Ferro N, Buitelaar P, Bogers T (2017) Semantic Representation and Enrichment of Information Retrieval Experimental Data. International Journal on Digital Libraries (IJDL) 18(2):145–172

Spärck Jones K, Bates RG (1977) Report on a design study for the 'ideal' information retrieval test collection. British Library Research and Development Report 5428, University Computer Laboratory, Cambridge

Spärck Jones K, van Rijsbergen CJ (1975) Report on the need for and provision of an 'ideal' information retrieval test collection. British Library Research and Development Report 5266, University Computer Laboratory, Cambridge

Voorhees EM (2002) The Philosophy of Information Retrieval Evaluation. In: Peters C, Braschler M, Gonzalo J, Kluck M (eds) Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross–Language Evaluation Forum (CLEF 2001) Revised Papers, Lecture Notes in Computer Science (LNCS) 2406, Springer, Heidelberg, Germany, pp 355–370

Voorhees EM (2007) TREC: Continuing Information Retrieval's Tradition of Experimentation. Communications of the ACM (CACM) 50(11):51–54

Voorhees EM, Harman DK (2005) TREC: Experiment and Evaluation in Information Retrieval. The MIT Press, MA, USA

W3C (2004) Resource Description Framework (RDF): Concepts and Abstract Syntax – W3C Recommendation 10 February 2004. https://www.w3.org/TR/rdf-concepts/

Zapilko B, Schaible J, Mayr P, Mathiak B (2013) TheSoz: A SKOS representation of the thesaurus for the social sciences. Semantic Web 4(3):257–263, DOI 10.3233/SW-2012-0081, URL http://dx.doi.org/10.3233/SW-2012-0081

Zobel J, Webber W, Sanderson M, Moffat A (2011) Principles for Robust Evaluation Infrastructure. In: Proc. Workshop on Data infrastructurEs for Supporting Information Retrieval Evaluation (DESIRE 2011), pp 3–6