Noname manuscript No.
(will be inserted by the editor)

Semantic Representation and Enrichment of Information Retrieval Experimental Data

Gianmaria Silvello · Georgeta Bordea · Nicola Ferro · Paul Buitelaar · Toine Bogers

PRE-PRINT: to appear in the International Journal on Digital Libraries

Received: date / Accepted: date

Abstract Experimental evaluation carried out in international large-scale campaigns is a fundamental pillar of the scientific and technological advancement of *Information Retrieval (IR)* systems. Such evaluation activities produce a large quantity of scientific and experimental data, which are the foundation for all the subsequent scientific production and development of new systems. In this work, we discuss how to semantically annotate and interlink this data, with the goal of enhancing their interpretation, sharing, and reuse. We discuss the underlying evaluation workflow and propose a *Resource Description Framework (RDF)* model for those workflow parts. We use expertise retrieval as a case study to demonstrate the benefits of our semantic representation approach. We employ this model as a means for exposing experimental data as *Linked Open Data (LOD)* on the Web and as a basis for enriching and automatically connecting this data with expertise topics and expert profiles.

In this context, a topic-centric approach for expert search is proposed, addressing the extraction of expertise topics, their semantic grounding with the LOD cloud, and their connection to IR experimental data. Several methods for expert profiling and expert finding are analysed and evaluated. Our results show that it is possible to construct expert profiles starting from automatically extracted expertise topics and that topic-centric approaches outperform state-of-the-art language modelling approaches for expert finding.

Keywords experimental data · expertise profiling · expert search · information retrieval evaluation · resource description framework · semantic enrichment

N. Ferro and G. Silvello
Department of Information Engineering
University of Padua, Padua, Italy
E-mail: {ferro, silvello}@dei.unipd.it

G. Bordea and P. Buitelaar
Insight Centre
National University of Ireland, Galway, Ireland
E-mail: {georgeta.bordea, paul.buitelaar}@insight-centre.org

T. Bogers
Department of Communication & Psychology
Aalborg University Copenhagen, Copenhagen, Denmark
E-mail: toine@hum.aau.dk

1 Introduction

The importance of research data is widely recognized across all scientific fields as this data constitutes a fundamental building block of science. Recently, a great deal of attention was dedicated to the nature of research data (Borgman, 2015) 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. In this context, the *Linked Open Data (LOD)* paradigm (Bizer et al., 2009; Heath and Bizer, 2011) has rapidly become the de-facto standard for publishing and enriching data. It allows for opening public data up in machine-readable formats ready for consumption, and for re-use and enrichment through semantic connections enabling new knowledge creation and discovery possibilities.

Several scientific fields started to expose research data as LOD on the Web. Relevant examples include applied life science research (Gray et al., 2014; Hersey et al., 2012), social sciences (Zapilko et al., 2013), linguistics (Di Buccio et al., 2013) and cultural heritage with the Europeana[1] LOD publishing effort (Isaac and Haslhofer, 2013), and the Library of Congress Linked

[1] `http://www.europeana.eu/`

Data Service[2], which "provides access to commonly found standards and vocabularies promulgated by the Library of Congress". Publishing houses are increasingly investing effort and money into exposing scientific publications metadata as LOD and into connecting publications with the underlying raw data. For instance, (i) Springer started an LOD project[3] for making the data about conference proceedings available and enriching their metadata with available data in the LOD cloud; (ii) Elsevier lauched the "Linked Data Repository"[4] with the aim to "store and retrieve content enhancements and other forms of semantic metadata about both Elsevier content"; and, (iii) in 2012 the Nature Publishing Group released a platform[5], which gives access to millions of LOD triples comprising "bibliographic metadata for all articles (and their references) from Nature Publishing Group and Palgrave Macmillan titles". Moreover, in the last years more than 100 data journals—whose aim is making research data effectively discoverable and reusable through data publications—has been proposed (Candela et al., 2015).

Paradoxically, in the field of *Information Retrieval (IR)*, where experimental evaluation based on shared data collections and experiments has always been central to the advancement of the field (Cleverdon, 1997; Harman, 2011), the LOD paradigm has not been adopted yet and no models or common ontologies for data sharing have been proposed. So despite the importance of data to IR, the field does not share any clear ways of exposing, enriching, and re-using experimental data as LOD with the research community. This impairs the reproducibility and generalization of IR experiments, which is rapidly becoming a key issue in the field. In 2011 the *ACM International Conference on Information and Knowledge Management* (CIKM) hosted the DESIRE workshop (Agosti et al., 2009) on data infrastructures for supporting IR evaluation with a specific focus on reproducibility. Since 2015 the *European Conference in IR (ECIR)* series has allocated a whole paper track on reproducibility; and, the 2015 edition of the *International ACM SIGIR Conference on Research and Development in IR* dedicated a specific workshop to the topic: *SIGIR Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR)* (Arguello et al., 2015). it is time to explore the possibility of (semi-) automatically maintaining and enriching the experimental data and of providing advanced services on top of them, as has been done in other scientific fields.

[2] `http://id.loc.gov/`
[3] `http://lod.springer.com/`
[4] `http://data.elsevier.com/`
[5] `http://data.nature.com/`

Therefore, the main objectives of this paper are to:

- define an RDF model of the scientific IR data with the aim of enhancing their discoverability and easing their connections with the scientific production related to and based on them;
- provide a methodology for automatically enriching the data by exploiting relevant external entities from the LOD cloud.

In particular, as far as the first objective is concerned, we define an RDF model (W3C, 2004a,b) for representing experimental data and exposing them as LOD on the Web. This will enable a 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 (Ferro and Silvello, 2014b).

As for the second objective, it builds upon the proposed RDF model and allows for automatically finding topics in the scientific literature exploiting the scientific IR data as well as connecting the dataset with other datasets in the LOD cloud. This augments the access points to the data as well as the potential for their interpretability and re-usability.

A positive collateral effect deriving from the pursuit of the second objective is the possibility of tackling the inherent complexity and heterogeneity of IR experimental data, which makes it difficult to find collaborators with an interest in a given topic or task, or to find all the experimental collections for a given topic. Identifying, measuring, and representing expertise has the potential to encourage interaction and collaboration—and ultimately knowledge creation—by constructing a web of connections between experts and the knowledge that they create. These connections allow individuals to access knowledge beyond their tightly-knit networks, where all members tend to have access to the same information. Additionally, expertise development is accelerated by providing valuable insight to outsiders and novice members of a community. In this way experimental data can be linked with underlying publications and associated people through extracted topics. The combination of experimental data with information extracted from related scientific narrative and semantic metadata help to enable a more meaningful interaction with them.

This paper is organised as follows. Section 2 discusses the workflow entailed by evaluation activities, presents an overview of the main challenges and existing solutions for modelling and managing experimental data in IR and describes state-of-the-art expert profiling and finding methodologies. Section 3 presents a concrete use case of our approach describing an RDF graph

of experimental data enriched with expertise topics, experts profiles, and links to external datasets. Section 4 tackles the first objective of this paper by presenting the parts of the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* conceptual model related to scientific production, experiments, semantic enrichment and expert profiling and the RDF model we defined. Section 6 builds on the presented RDF model and tackles the second and third objective of the paper by defining the enrichment process of experimental data based on the publications related to evaluation campaigns and background knowledge available on the LOD cloud. In Section 7, we discuss several experiments for assessing the effectiveness and the semantic grounding of expertise topics, expert finding, and expert profiling. We conclude this paper by presenting some final remarks and future work in Section 8.

2 Background

2.1 Experimental evaluation in IR

IR is concerned with complex systems delivering a variety of key applications to industry and society: Web search engines (Croft et al., 2009), (bio)medical search (Müller, 2013), enterprise search (Burnett et al., 2006), intellectual property and patent search (Lupu and Hanbury, 2013), expertise retrieval systems (Balog et al., 2012), and many others.

Therefore, designing and developing these faceted and complex systems are quite challenging activities and, since the inception of the field, they have been accompanied by thorough experimental evaluation methodologies, in order to be able to measure the achieved performance, to assess the impact of alternatives and new ideas, and to ensure the levels of effectiveness needed to meet user expectations.

Experimental evaluation is a demanding activity in terms of effort and required resources, that benefits from using shared datasets, which allow for repeatability of the experiments and comparison among state-of-the-art approaches (Ferro and Silvello, 2015, 2016; Kharazmi et al., 2016). Therefore, over the last 20 years, experimental evaluation is carried out in large-scale evaluation campaigns at the international level, such as the *Text REtrieval Conference (TREC)*[6] (Harman and Voorhees, 2005) in the US, the *Conference and Labs of the Evaluation Forum (CLEF)*[7] (Ferro, 2014) in Europe, or the *NII Testbeds and Community for Information access Research (NTCIR)*[8] in Japan and Asia.

Over the years, these evaluation activities have provided sizable results and improvements in various key areas, such as indexing techniques, relevance feedback, multilingual search, results merging, and so on. For example, before CLEF started in 2000, the best bilingual information access systems performed about 45%-50% as well as the corresponding best monolingual systems (Harman et al., 2001), further limited to resource-rich languages such as English, French, and German. After ten years of CLEF, the best bilingual systems went up to about 85%-95% of the best monolingual ones (Agirre et al., 2009; Ferro and Silvello, 2014a) for most language pairs. Over the years, these initiatives have resulted in massive amounts of scientific data, comprising shared datasets, experimental results, performance measures, descriptive statistics and statistical analyses about them, which provided the foundations for the subsequent scientific and technological development. Consequently, experimental data as well as evaluation campaigns have a high scientific and economical value. TREC estimated that for every $1 that the *National Institute of Standards and Technology (NIST)* and its partners have invested in TREC, at least $3.35 to $5.07 in benefits accrued to researchers and industry which, for an overall investment in TREC of around 30 million dollars in 20 years means producing between 90 and 150 million dollars of benefits (Rowe et al., 2010).

Experimental evaluation (Harman, 2011; Robertson, 2008) is a very strong and long-lived tradition in the IR field and dates back to the late 1950s/early 1960s. It is based on the Cranfield methodology (Cleverdon, 1997) which makes use of shared experimental collections in order to create comparable experiments and evaluate the performance of different information access systems. An experimental collection is a triple composed of: (i) a set of documents, also called collection of documents, which is representative of the domain of interest in terms of both kinds of documents and number of documents; (ii) a set of topics, which simulate actual user information needs and are used by IR systems to produce the actual queries to be answered; and, (iii) the ground-truth or the set of relevance judgments, i.e., a set of 'correct' answers, where for each topic the documents which are relevant for that topic, are determined.

In Figure 1 we can see the main phases of the IR experimental evaluation workflow, where an increased attention for the knowledge process entailed by evaluation campaigns is required. Indeed, the complexity of the tasks and the interactions to be studied and eval-

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

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

[8] `http://research.nii.ac.jp/ntcir/`

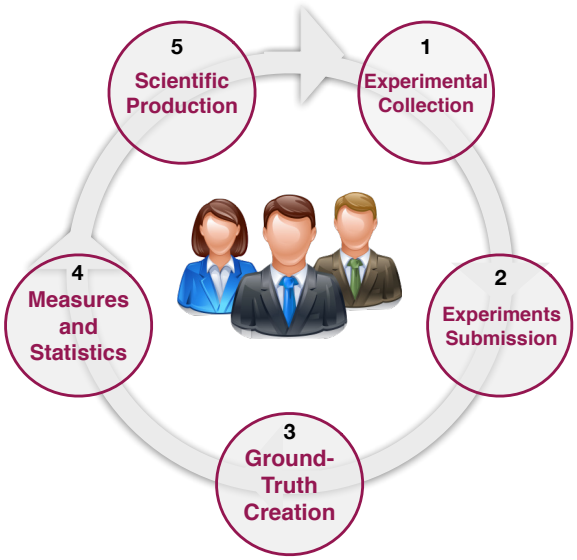


Fig. 1 The typical IR experimental evaluation workflow and the data produced

uated produce valuable scientific data, which need to be properly managed, curated, enriched, and accessed. Further, the information and knowledge derived from them needs to be appropriately treated and managed, as well as the cooperation, communication, discussion, and exchange of ideas among researchers in the field. In this perspective, the information space entailed by evaluation campaigns can be considered in the light of the *Data, Information, Knowledge, Wisdom (DIKW)* hierarchy (Ackoff, 1989; Fricke, 2009; Rowley, 2007; Zeleny, 1987), used as a model to organize the produced information resources (Dussin and Ferro, 2009).

The first step regards the creation of the experimental collection and is composed of the acquisition and preparation of the documents and the creation of topics from which a set of queries is generated. In the second step, the participants in the evaluation campaign have everything they need to run experiments and test their systems. An experiment is the output of an IR system, which usually consists of a set of ranked lists of documents—one list per topic. Both the experimental collections and the experiments can be regarded as *data*, since they are raw and basic elements, which have little meaning by themselves and no significance beyond their existence.

In the third step, the gathered experiments are used by the campaign organizers to create the ground-truth, typically adopting some appropriate sampling technique to select a subset of the dataset for each topic. In this phase, assessors decide whether or not an object is relevant for a given topic. Relevance judgments are raw data belonging to the experimental collection, but at the same time they represent human-added information connecting documents to topics of an experiment. The triple of documents, topics, and relevance judgments is then used to compute performance measures for each experiment. Both relevance judgements and performance measures can be considered as *information*, since they associate meaning with the data through some kind of relational connection and are the result of computations on and processing of the data.

Afterwards, measurements are used to produce descriptive statistics and conduct statistical tests about the behavior of one or more systems, which represents *knowledge*, as these tests are built upon the performance measurements and used to make decisions and take further action on future scientific work.

Finally, the last step is 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. This phase corresponds to the *wisdom* in the DIKW hierarchy. Furthermore, this phase also embraces external actors who were not originally involved in the evaluation campaign. Indeed, the data employed in the evaluation workflow (i.e., documents, topics, and relevant judgments) as well as the data produced (i.e., experiments, measures and statistics, and reports) are usually made freely available to the scientific community, which exploits them to produce new knowledge in the form of scientific papers. Scientific production is central to the evaluation workflow, because it involves all the data used and produced in the process, all the actors who participated to the campaign, and external actors who may exploit and elaborate upon the data.

In this article, we focus in particular on this last step, providing an RDF model of the resources involved in the scientific production and management and leveraging it as the starting point for extracting expert profiles and topics, which are used both to semantically enrich the underlying scientific data and to link them to other data sources in the LOD cloud.

2.2 Modeling and managing IR experimental data

A crucial question in IR, common to other research fields as well, is how to ensure the best exploitation and interpretation of the valuable scientific data employed and produced by experimental evaluation, possibly over large time spans. For example, the importance of describing and annotating scientific datasets

is discussed by Bowers (2012), who notes that this is an essential step for their interpretation, sharing, and reuse. However, this question is often left unanswered in the IR field since researchers are more interested in developing new algorithms and methods for innovative systems than modeling and managing their experimental data (Agosti et al., 2007a,b). As a consequence, we have started an effort 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 (Agosti et al., 2012b), as well as 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). DIRECT enables the evaluation workflow described in the previous section, manages the scientific data produced during a large-scale evaluation campaign, as well as supports the archiving, access, citation, dissemination, interaction, and sharing of the experimental results (Agosti et al., 2012a; Agosti and Ferro, 2009; Angelini et al., 2014, 2016; Ferro et al., 2011). To our best 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.

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 accordingly to the necessities and the requirements provided by the community (Agosti et al., 2012b). At the time of writing, DIRECT handles about 35 million documents, 14 thousand topics, around 4 million relevance judgments, 5 thousand experiments and 20 million measures. This data has been used by about 1,500 researchers from more than 70 countries worldwide. Overall, DIRECT counts around 650 visitors who accessed and downloaded the data thus proving that the DIRECT system is well-suited to address most of the community requirements for what it is concerned with the access and use of IR experimental data.

There are other projects with similar goals but with a narrower scope. One such project is the *Open Relevance Project (ORP)*[9] 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 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 does it take into account the scientific production connected to the experimental data, which is vital for the enrichment of the data themselves as well as for, for instance, the definition of expert profiles.

Another relevant project is EvaluatIR.org[10] (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 (Testbed for Information Retrieval Algorithms) 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. We can also mention some other efforts made by the community to provide toolkits to support the different phases of Machine Learning/IR experiments such as WEKA[11] and the SimDL framework (Leidig, 2012) which integrates a digital library with simulation infrastructures to provide automated support for research components. Although these services are relevant to the field, they are not directly related to DIRECT which aim is to model and manage the whole evaluation process in IR and to provide access to the evaluation products rather than to propose new evaluation methodologies or to provide researchers with new tools for carrying out their activities.

Thorough modeling and managing of experimental data and the related scientific publications is fundamental for creating new knowledge on top of this data; to this purpose, DIRECT and the modeled evaluation workflow are the starting point we consider for exposing experimental collections and related scientific contributions as LOD on the Web. To the best of our knowledge a semantic model for representing IR experimental data has been proposed here for the first time. Furthermore, as a relevant outcome of this approach we also show how it is possible to exploit scientific contributions for enriching experimental data as well as for automatically

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

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

[11] http://www.cs.waikato.ac.nz/ml/weka/

defining expert profiles on a series of identified scientific topics.

2.3 Expert finding and profiling

Expert finding is the task of locating individuals knowledgeable about a specific topic, while *expert profiling* is the task of constructing a brief overview of the expertise topics of a person. So far, these tasks have been deemed interesting mainly for their application in enterprise settings. However, scientific communities could benefit from such tasks and tools as well, as they enable and strengthen collaboration. In an academic setting, existing work on expert finding focused on the task of finding qualified reviewers to assess the quality of research submissions (Mimno and McCallum, 2007; Rodriguez and Bollen, 2008). In this work, we consider its applications for dissemination and sharing of experimental results in IR.

Initial solutions for expert finding were developed under the area of competency management (Draganidis and Metzas, 2006). These approaches are based on manual construction and querying of databases about knowledge and skills of an organization's workforce, placing the burden and responsibility of maintaining them on the employees themselves (Maybury, 2006). A disadvantage of this approach is that because the information about experts and expertise is highly dynamic, considerable efforts are required to keep competency databases up-to-date. This prompted a shift to automated expert finding techniques that support a more natural expertise location process (Campbell et al., 2003).

Expert finding can be modelled as an information retrieval task using queries provided by users to perform a full-text search for experts instead of documents. The goal of the search is to create a ranking of people who are experts on a given topic, instead of ranking relevant documents. A lot of ground was covered in terms of evaluating expert search systems by the organisation of three consecutive enterprise tracks by TREC (Bailey et al., 2007), that provided common ground for evaluating different systems and approaches. In this context, the expert finding task is modelled using statistical language modelling (Balog et al., 2006; Petkova and Croft, 2006) or data fusion techniques (Macdonald and Ounis, 2006).

The importance of expert profiling when developing solutions for expert search is discussed in (Balog and de Rijke, 2007), without addressing the problem of discovery and identification of expertise topics. The authors assume that a controlled vocabulary of terms is readily available for the considered domain. Currently, such a resource is not available in our application setting, so we therefore propose an automatic solution for the extraction of expertise topics by adapting existing term extraction and keyphrase extraction approaches.

An extensive analysis of expert profiling is presented in (Serdyukov et al., 2011), where the language model proposed by (Balog and de Rijke, 2007) is included as one of the features in a machine learning approach. Other features include a more simple binary model of relevance and the frequency of an expertise topic in expert profiles from the training set. Expertise topics, called tags in this work, are assumed to be known in advance, similar to (Balog and de Rijke, 2007), and are collected through self-assessment. An important observation is that the quality of expertise topics is more important than the relevance to a particular person. In their experiments, the most important feature with respect to its performance contribution is the frequency of the expertise topic, a feature that is independent of the particular employee.

We build on this work by using a quality-related measure of expertise topics together with relevance-based measures for expert profiling. Similar to competency management approaches, an intermediate conceptual level between documents and experts is introduced, avoiding their limitations, such as manual gathering of data and quickly outdated profiles through automatic extraction of expertise topics.

3 Use Case: Discover, Understand and Re-use IR Experimental Data

As previously discussed, in order to allow for a better understanding and re-use of experimental IR data and to increase their potential, visibility and discoverability on the Web, we start from a well-established and comprehensive conceptual model of experimental data—realized by the DIRECT system—and we provide a mapping towards an RDF model letting us to expose these data as LOD on the Web. This is the first step towards improving the possibility of discovering experimental data and enriching them to augment their understandability and re-usability. Following this line of work, we adopt an automatic approach for extracting expertise topics from the contributions connected to the experimental data and then use them for enriching the contributions themselves and their authors, connecting with the LOD cloud, and defining expert profiles.

In this section, we discuss an example of the outcomes of the semantic modeling and automatic enrichment processes applied to the use case of discovering, understanding and re-using the experimental data. Figure 2 shows an RDF graph, which provides a visual rep-

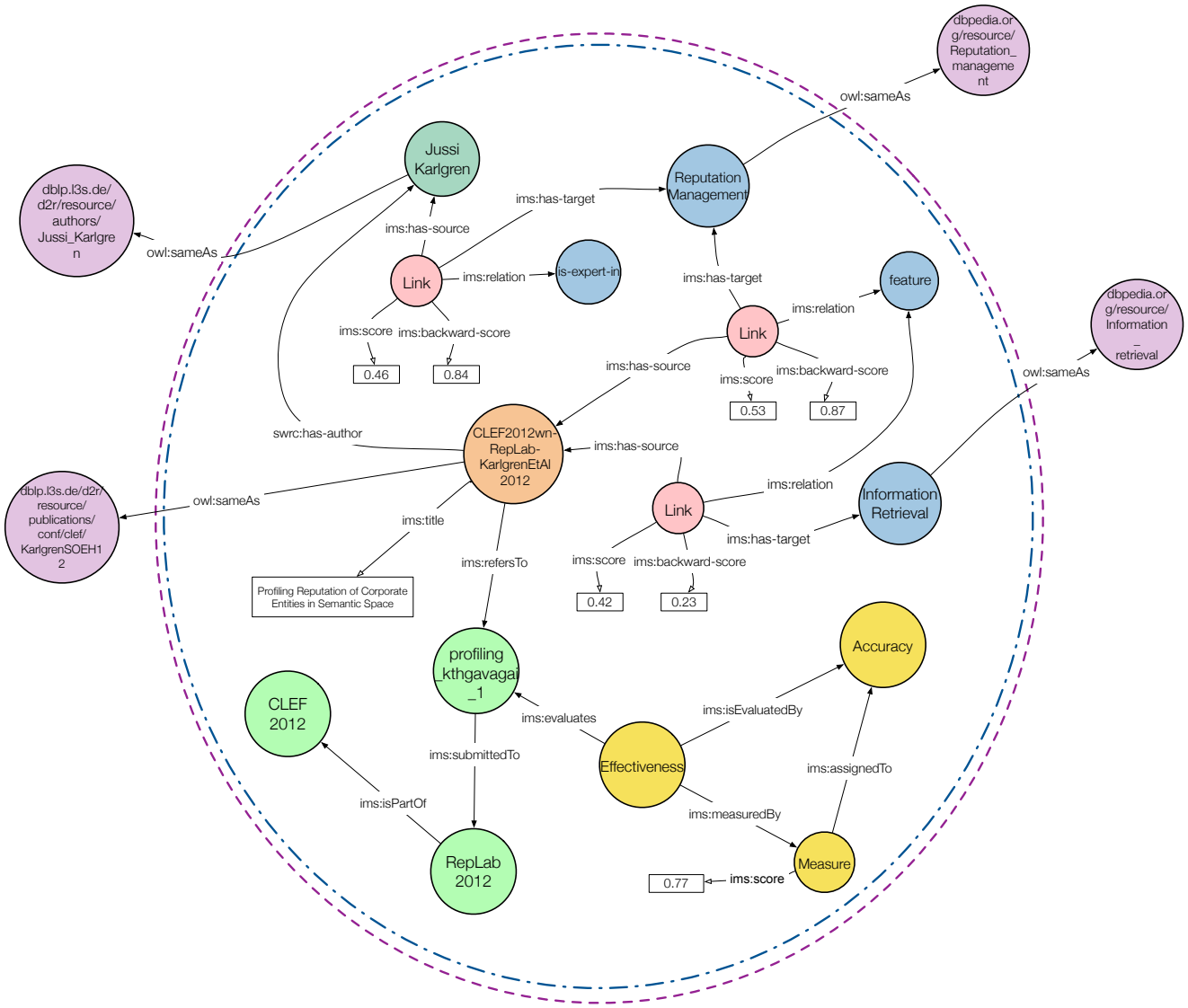


Fig. 2 An example of RDF graph showing how expertise topics and expert profiles are used for enriching IR experimental data. The colors assigned to the classes identify the different areas of the DIRECT model which are described in Section 4.

Table 1 Colors used in the RDF graphs and DIRECT areas.

Color	DIRECT Area
Dark green	Management area
Light green	Evaluation activity area
Yellow	Measurement area
Light blue	Concept area
Orange	Contribution area
Blue	Expert profile/topic area
Purple	External classes (LOD Cloud)

resentation of how the experimental data are enriched. In particular, we can see the relationship between a contribution and an author enriched by expertise topics, expert profiles and connections to the LOD cloud. As reported in Table 3, the different classes are associated with different colors in order to identify the different areas of the RDF model they belong to. For instance, the class representing the `User` is colored in dark green as the classes of the management area described in Section 4; the classes regarding expertise topics and expert profile enrichment are colored in blue, the classes related to the measures within the experiments area are colored in yellow, the classes related to the evaluation activities (tasks, tracks and campaign) are colored in light green and the `Contribution` class is colored in orange.

In this instance, the author (*Jussi Karlgren*) and the contribution (*KarlgrenEtAl-CLEF2012*) are data derived from the evaluation workflow, whereas all the other information are automatically determined by the enrichment process. The adopted methodology for expertise topics extraction determined two main topics, “reputation management” and “information retrieval”,

which are related to the *KarlgrenEtAl-CLEF2012* contribution. These topics are connected to the contribution by instantiating the RDF model shown in Figure 5 and discussed below, by using the `Link` class which acts as an RDF blank node[12]. 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 (normalized in the interval $[0, 1]$) give a measure of how much this contribution is about a specific topic and we can see that in this case it is concerned a bit more with reputation management than with information retrieval. Furthermore, the backward-score gives us additional information by measuring how much a contribution is authoritative with respect to a scientific topic. In Figure 2, we can see that *KarlgrenEtAl-CLEF2012* is authoritative for reputation management (backward-score of 0.87), whereas it is not a very important reference for information retrieval (backward-score of 0.23). Summing up, we can say that if we consider the relation between a contribution and an expertise topic, the score indicates the pertinence of the expertise topic within the contribution; whereas the backward score indicates the pertinence of the contribution within the expertise topic. The higher the backward score, the more pertinent is the contribution for the given topic.

This information is confirmed by the expert profile data; indeed, looking at the upper-left part of Figure 2, the author *Jussi Karlgren* is considered "an expert in" reputation management (backward-score of 0.84), even if it is not his main field of expertise (score of 0.46).

All of this automatically extracted information enriches the experimental data enabling for a higher degree of re-usability and understandability of the data themselves. In this use case, we can see that the expertise topics are connected via an `owl:sameAs` property to external resources belonging to the DBPedia[13] linked open dataset. These connections are automatically defined via the semantic grounding methodology described below and enable the experimental data to be easily discovered on the Web. In the same way, authors and contributions are connected to the DBLP[14] linked open dataset.

In Figure 2 we can see how the contribution (*KarlgrenEtAl-CLEF2012*) is related to the experiment (*profiling_kthgavagai_1*) on which it is based. This experiment was submitted to the *RepLab 2012* of the evaluation campaign *CLEF 2012*. It is worthwhile to highlight that each evaluation campaign in DIRECT is defined by the name of the campaign (CLEF) and the year it took place (e.g., 2012 in this instance); each evaluation campaign is composed of one or more tasks identified by a name (e.g., RepLab 2012) and the experiments are treated as submissions to the tasks. Each experiment is described by a contribution which reports the main information about the research group which conducted the experiment, the system they adopted, developed and any other useful detail about the experiment.

We can see that most of the reported information are directly related to the contribution and they allow us to explicitly connect the research data with the scientific publications based on them. Furthermore, the experiment is evaluated from the "effectiveness" point of view by using the "accuracy" measurement which has 0.77 score. Retaining and exposing this information as LOD on the Web allow us to explicitly connect the results of the evaluation activities to the claims reported by the contributions.

4 Data Modeling for Enrichment and Expert Profiling

The detection of scientific topics related to the data produced by the experimental evaluation and the creation of expert profiles mainly concerns three areas covered by the evaluation workflow, which we call the "resource management area" (Figure 3), the "experiment area" (Figure 4)), and the "scientific production area" (Figure 7(b)). As described above, DIRECT covers all of these workflow aspects, which leads to a rather complex system, the presentation of which is out of the scope for this paper (Agosti et al., 2012b). Nevertheless, the conceptual model of the DIRECT resource management and scientific production areas has been mapped into an RDF model and adopted for enriching and sharing the data produced by the evaluation activities.

Within this model we consider a `Resource` as a generic class sharing the same meaning of resource in RDF (W3C, 2004b) where "*all things described by RDF are called resources. [A resource is] the class of everything.*" In DIRECT a `Resource` represents the class of everything that exists in IR experimental evaluation.

The resource management area models the more general and coarse-grained resources involved in the evaluation workflow—i.e., users, groups, roles, namespaces, and concepts—and the relationships among them. Furthermore, it handles the provenance of the data. All the classes of this area are defined as subclasses of the general `Resource` class and they are represented

[12] Blank nodes do not have identifiers in the RDF abstract syntax. The blank node identifiers have local scope and are purely an artifact of the serialization. Refer to `http://www.w3.org/TR/rdf11-concepts/#section-blank-nodes` for more details on blank nodes.

[13] `http://www.dbpedia.org/`

[14] `http://dblp.l3s.de/`

Table 2 RDF namespaces and prefixes of the vocabularies adopted in DIRECT for the Resource Management, Experiment, and Scientific Production areas.

Prefix	RDF namespace	Description
aktors	http://www.aktors.org/ontology/portal#	Advanced Knowledge Technology reference ontology
basic	http://def.seegrid.csiro.au/isotc211/iso19103/2005/basic	OWL representation of ISO 19103
bibo	http://purl.org/ontology/bibo/	Bibliographic ontology
dcterms	http://purl.org/dc/terms/	Dublin Core terms
foaf	http://xmlns.com/foaf/0.1/	Friend of a friend
ims	http://ims.dei.unipd.it/data/rdf/	DIRECT vocabulary terms
org	http://www.w3.org/ns/org	Core organization ontology
owl	http://www.w3.org/2002/07/owl#	OWL vocabulary terms
prov	http://www.w3.org/ns/prov#	The ontology supporting the interchange of provenance on the web
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	RDF vocabulary terms
rdfs	http://www.w3.org/2000/01/rdf-schema#	RDF Schema
salt	http://salt.semanticauthoring.org/ontologies/sdo#	SALT Document Ontology
schema	https://schema.org/	Schema.org promotes schemas for structured data on the Internet, on web pages, in email messages, and beyond.
skos	http://www.w3.org/2009/08/skos-reference/skos.html	SKOS Simple Knowledge Organization System
swco	http://data.semanticweb.org/ns/swc/swc_2009-05-09.html#	Semantic Web Conference Ontology
stato	http://stato-ontology.org/	A general-purpose STATistics Ontology
swpo	http://sw-portal.deri.org/ontologies/swportal#	Semantic Web Portal Ontology
swrc	http://swrc.ontoware.org/ontology#	Semantic Web for Research Communities ontology
vann	http://purl.org/vocab/vann/	Vocabulary for annotating descriptions of vocabularies
vcard	http://www.w3.org/2006/vcard/ns#	vCard electronic business card profile defined by RFC 2426
xsd	http://www.w3.org/2001/XMLSchema#	XML Schema

in Figure 3 along with the properties connecting them; for sake of readability we omitted from the figure the datatype properties which are reported in Table 3.

In Figure 3 it is also possible to see how the classes of the management area are related to other vocabularies (reported in Table 2) in the LOD cloud throughout the `schema:isSimilarTo` and the `owl:sameAs` properties; the first one is used to establish a semantic relation between two similar concepts, whereas the second one is used to establish a formal equality between them. These relationships open up new entry points to the DIRECT dataset and new reasoning possibilities over the experimental data.

The `User` class, related to the `Agent` class of the `foaf` vocabulary, represents the actors involved in the evaluation activities such as researchers conducting experiments, organizers of a campaign, assessors, data scientists, and authors of scientific contributions. Furthermore, the `User` class as well as the `foaf:Agent` class may enclose also non-human agents such as software libraries. The function of a user in the evaluation workflow is defined by the `Role` class and the users can be grouped together via the `Group` class. A user can play none, one, or more roles: for instance, a user can be both an assessor as well as a researcher submitting experiments, i.e., a participant to the campaign. On the other hand, there are roles played by more then one user; for instance, a campaign can have one or more participants, e.g., the researchers that are carrying out the experiments for writing a paper. A group is a resource arranging together users with some common characteristics; for instance, there could be a group formed by all the users belonging to a specific research group or an ad-hoc group created just for one project or collaboration. The `Role` class is related to role class in the `org` and `swco` vocabularies, whereas `Group` is related to the corresponding `foaf` one.

The `Namespace` class refers to a logical grouping of resources, allows the disambiguation of homonyms, and is related to the namespace class of the `vann` vocabulary (see Table 2). The use of namespaces in DIRECT is different from the namespace mechanism in *eXtensible Markup Language (XML)* and RDF which is used "to associate the schema with the properties in the assertions"[15]; indeed, in DIRECT, namespaces are used to organize resources of the same kind but com-

[15] `http://www.w3.org/TR/WD-rdf-syntax-971002/`

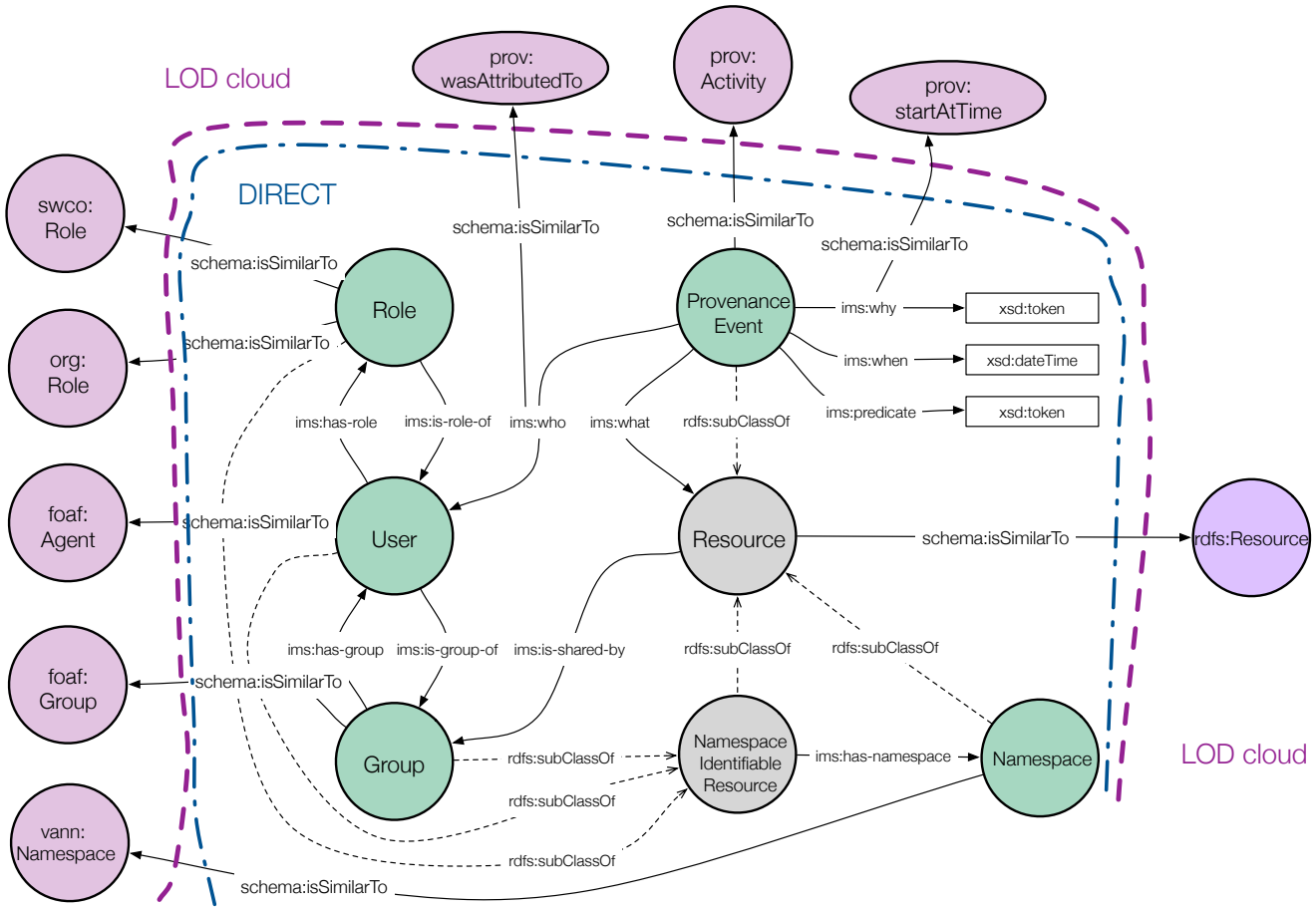


Fig. 3 The Resource Management area classes and properties.

ing from different domains. For instance, in the context of experimental evaluation, we could host data coming from CLEF and TREC and assign them to two different namespaces, to clearly separate them. In the RDF model of DIRECT along with the general `Resource` we described above, there is another general class called `Namespace Identifiable Resource` as we can see in Figure 3; this is a subclass of `Resource` always associated to a namespace. Thus, in the RDF model of DIRECT there are two kinds of general resources, the first which has no namespace and the second which has one. In Figure 3 we can see that `User`, `Group` and `Role` have a namespace, whereas the `Namespace` itself and `Provenance-Event` have not. The resources with a namespace are those that can be logically grouped or disambiguated by using some of their inner characteristics such as affiliation for the users or venue for the contributions; the resources without a namespace are those that do not need to be grouped or disambiguated such as the provenance events which are handled internally to the system and thus do not need to be disambiguated (there are not two events with the same name or identifier) or logically grouped.

Finally, the `Provenance-Event` class keeps track of the full lineage of each resource managed by DIRECT since its first creation, allowing users with adequate access permissions to reconstruct its full history and modifications over time. This class is related to the `Activity` class in the `prov` ontology (see Table 2). As shown in Figure 3, `Provenance-Event` is composed of two object properties and three datatype properties, where:

- **who**, is the property associating the provenance event with the user who caused the event;
- **what**, is the property associating the provenance event with the specific resource originated by the event—note that every resource in the model can be related to a provenance event;
- **when**, is the datatype property associating the provenance event with the timestamp at which the event occurred;
- **why**, is the datatype property associating the provenance event with the motivation that originated the event, i.e., the operation performed by the system that led to a modification of the resource;

Table 3 Main datatype properties of the resource management and contribution area classes reported in Figures 3 and 5. `Namespace` `Identifiable Resource`, `Concept`, `Group`, and `Role` are not reported because they have no additional datatype properties w.r.t. `Resource`. "ims" (as per Information Management Systems) is the prefix for `http://ims.dei.unipd.it/data/rdf/` pointing to DIRECT vocabulary terms.

Class	OWL Datatype Properties	xsd:datatype
`Contribution`	ims:affiliation	xs:string
	ims:title	xs:string
	ims:pages	xs:string
	ims:additional-information	xs:string
	ims:year	xs:gYear
	ims:link	xs:anyURI
	ims:copyrighted	xs:boolean
`Link`	ims:score	xs:double
	ims:backward-score	xs:double
	ims:frequency	xs:positiveInteger
`Namespace`	ims:prefix	xs:string
`Provenance-Event`	ims:when	xs:dateTime
	ims:why	xs:token
	ims:predicate	xs:token
`Resource`	ims:identifier	xs:token
	ims:created	xs:dateTime
	ims:last-modified	xs:dateTime
	ims:description	xs:string
	ims:name	xs:string
	ims:content	xs:string
	ims:content-transfer-encoding	xs:token
`User`	ims:password	xs:string
	ims:first-name	xs:string
	ims:last-name	xs:string
	ims:affiliation	xs:string
	ims:e-mail	xs:string
	ims:birth-date	xs:date
	ims:gender	xs:token
	ims:address	xs:string
	ims:city	xs:string
	ims:state	xs:string
	ims:zip	xs:string
	ims:phone	xs:string
	ims:facsimile	xs:string
	ims:mobile	xs:string
	ims:voip-caller-id	xs:token
	ims:homepage	xs:anyURI

– **predicate**, is the datatype property associating the provenance event with the action carried out in the event, i.e., CREATED, READ, or DELETED.

Modeling provenance is key to the definition of expert profiles and topic extraction, because it is a means for controlling the quality and integrity of the data produced by the evaluation workflow. As we discussed above, the data produced by experimental evaluation are not only raw data, but they are the product of a series of transformations, which involve inputs from scientists and experts of the field. Retaining what was done with the data is crucial if we want to verify the quality, to reproduce the experiments (Buneman, 2013), to share (Borgman, 2012), and to cite (Silvello, 2015) the raw data or their elaborations. Moreover, this data is used for scientific production, which in turn are exploited for expert profiling—two activities that must rely on high quality data. The `Provenance-Event` class allows us to record the five aspects (i.e., who, what, when, why and predicate) required for keeping the lineage of data (Cheney et al., 2009) and, consequently, the reliability of the information we extract and infer from this data.

In Figure 4 we can see the classes and properties of the experiment area and the relationships of these classes with external classes in the LOD cloud. This area can be divided in two main parts, the first comprehending the `Run`, `Track` and `Evaluation Activity` classes modeling the experiments (i.e., *runs* in the experimental evaluation campaigns vocabulary) and the

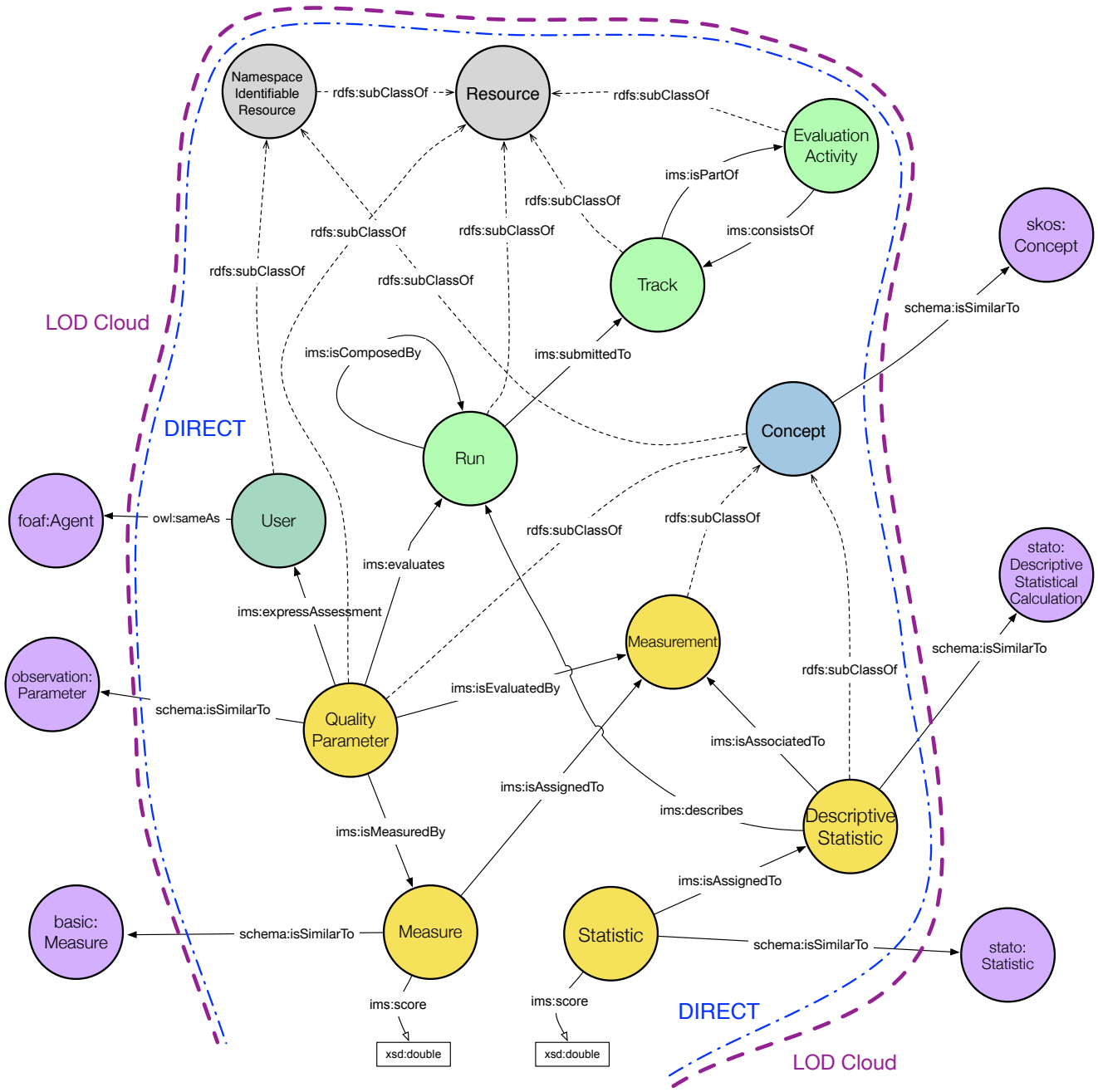


Fig. 4 The Experiment area classes and properties.

second 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. We can see that the class `Run` has a recursive property called `isComposedBy` which allows for defining an experiment as a composition of smaller experiments; such experiments are quite common in a typical IR scenario where a run is composed of one ranked list of objects for each topic in an experimental collection. In this case, a ranked list for a given topic represents an experiment and the run, which is the union of the ranked lists of all topics, represents a group of experiments which in DIRECT is defined as an "experiment of experiments". The class `Run` allows us to handle a run as a whole or as single parts corresponding to each single topic in the collection.

The second part allows us to model the measurements and the descriptive statistics calculated from the runs. It is built following the model of quality for *Digital Library (DL)* defined by the DELOS Reference Model (Can-

dela et al., 2007) which is a high-level conceptual framework that aims at capturing significant entities and their relationships with 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 as detailed in (Agosti et al., 2016). The quality domain in the DELOS Reference model takes into account the general definition of quality provided by *International Organization for Standardization (ISO)* which defines quality as "*the degree to which a set of inherent characteristics fulfils requirements*" (ISO 9000, 2005), where requirements are needs or expectations that are stated, generally implied or obligatory while characteristics are distinguishing features of a product, process, or system (Agosti et al., 2007c, 2016). A `Quality Parameter` is a `Resource` that indicates, or is linked to, performance or fulfillment 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 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, 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 (e.g., "effectiveness" or "efficiency") and resemble the notion of quality dimension in (Batini and Scannapieco, 2006). They express the assessment of an `User` about the `Resource` under examination. They can be evaluated according to different `Measurement`s (e.g., "accuracy" as in Figure 2 or commonly used measurements such as "precision" or "recall"), 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 IR system is its effectiveness, which is its capability of answering user information needs by retrieving relevant items. This `Quality Parameter` can be evaluated according to many different `Measurement`s, such as precision and recall (Salton and McGill, 1983): precision evaluates effectiveness in the sense of the ability of the system to reject useless items, whereas recall evaluates effectiveness in the sense of the ability of the system to retrieve useful items. The actual values for precision and recall are `Measure`s and are usually computed using standard tools, such as `trec_eval`[16], which are `Users`, but in this case not human.

Furthermore, the `Descriptive Statistic` class models the possibility of associating statistical analyses with 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.

Lastly, another important class of the model is `Concept` which is defined as an idea or notion, a unit of thought. It is used to define the type of relationships in a semantic environment or to create a vocabulary (e.g., contribution types) and it resembles the idea of concept introduced by the *Simple Knowledge Organization System (SKOS)* (W3C, 2009a,b) to which it is related via the `schema:isSimilarTo` property. `Concept` is a subclass of `Namespace Identifiable Resource` and thus its instances are always associated with a namespace.

In DIRECT every vocabulary we create or import is handled via the `Concept` class. As an example, let us consider the term "Book" taken from the "Advanced Knowledge Technology reference ontology" which has `http://www.aktors.org /ontology/portal#` as *Uniform Resource Identifier (URI)* and prefix "aktors" (see Table 2). In Figure 6 we can see how the model reported in Figures 3 and 5 is instantiated for representing this term. We can see that the URI of the "aktors" ontology is retained by the URI of the instance of the `Namespace` class (which in the figure is renamed as "aktors URI" for convenience), whereas the prefix is represented by the datatype property `ims:prefix`. In Table 2 we report the vocabularies adopted in DIRECT for the resource management and scientific production areas.

In Figure 5 we can see the classes and the properties of the scientific production area; also in this case we show the relationships between the classes of this area and the external classes in the LOD cloud. This area of the RDF model is central for the expert profiling activity because it handles scientific contributions, their relations with scientists and authors, and the scientific topics that can be extracted from them. Figure 5 reports three main classes which are `Concept`, `Contribution` and `Link`.

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

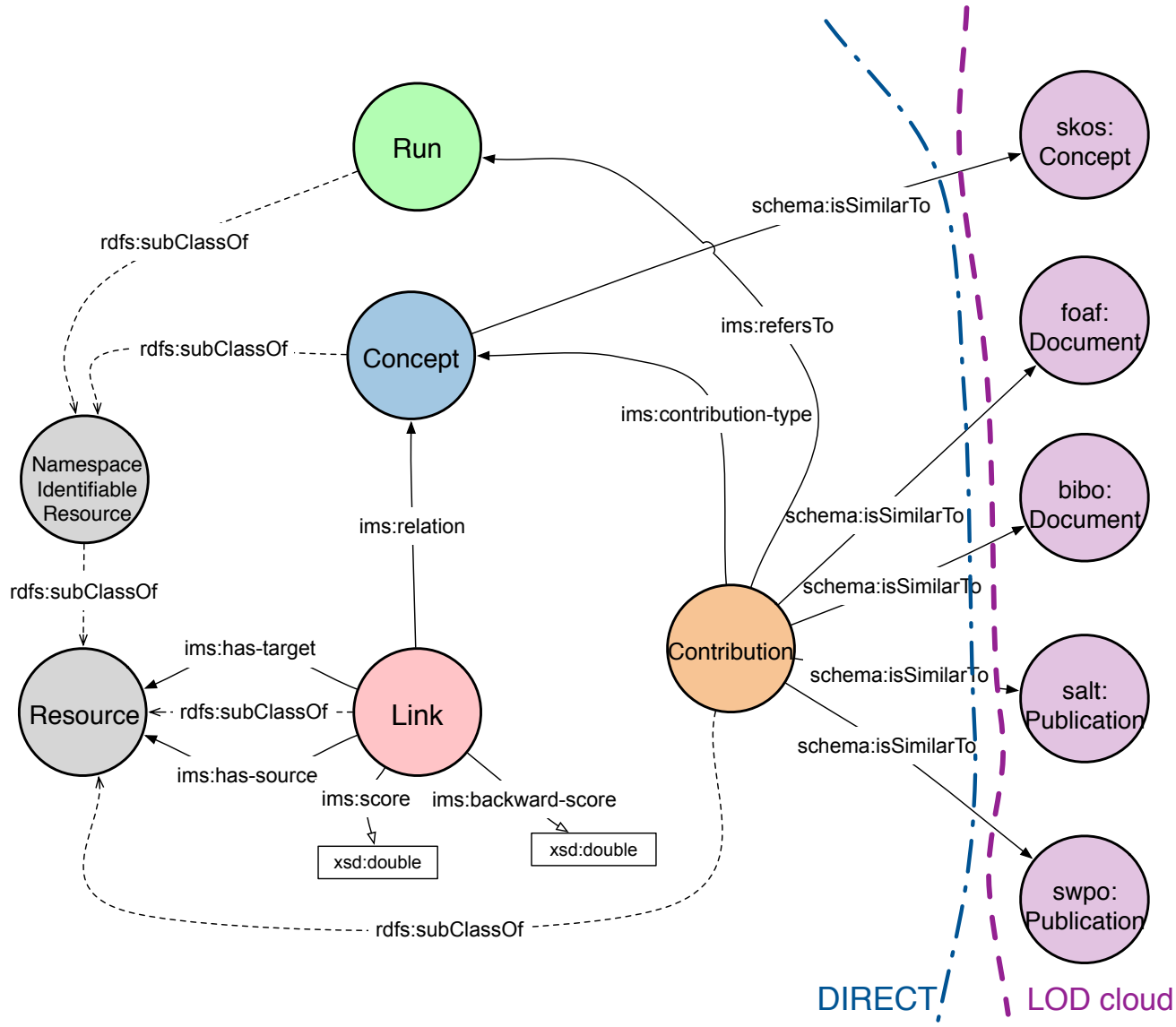


Fig. 5 The Scientific Production area classes and properties.

The `Contribution` class represents every publication concerning the scientific production phase of the evaluation workflow. We can see that it is related to `Concept` via the `ims:contribution-type` property which can be instantiated as shown in Figure 6.

The `Contribution` class is related to four similar classes from external datasets: `Document` from the `bibo` and `foaf` vocabularies and `Publication` from the `salt` and `swpo` ones.

The `Link` class connects two resources via the `ims:has-source` and `ims:has-target` properties with a typed relationship realized throughout a concept connected to the link via the `ims:relation` property. This allows us to create typed relationships between two generic resources involved in the evaluation workflow. We can instantiate the graph in Figure 5 in several ways; a first example is shown in the right part of Figure 6, where we represent two terms (i.e., "Publication" and "Book") belonging to a vocabulary. This very example can be extended by representing a taxonomy of terms belonging to one or more vocabularies. In Figure 6 we can see how the "Book" term can be related throughout an "is-a" relation to the more general term "Publication". So, in this case `Link` is instantiated by a generic "Link" resource, which relates two concepts, i.e., "Book" and "Publication", via the `ims:has-source` and `ims:has-target` datatype properties. The datatype property `ims:relation` allows us to define the type of the relationship—"is-a" in this case—between the two associated concepts.

The concept "Book" is associated with the instance "contributionX" of `Contribution` by means of the `ims:`

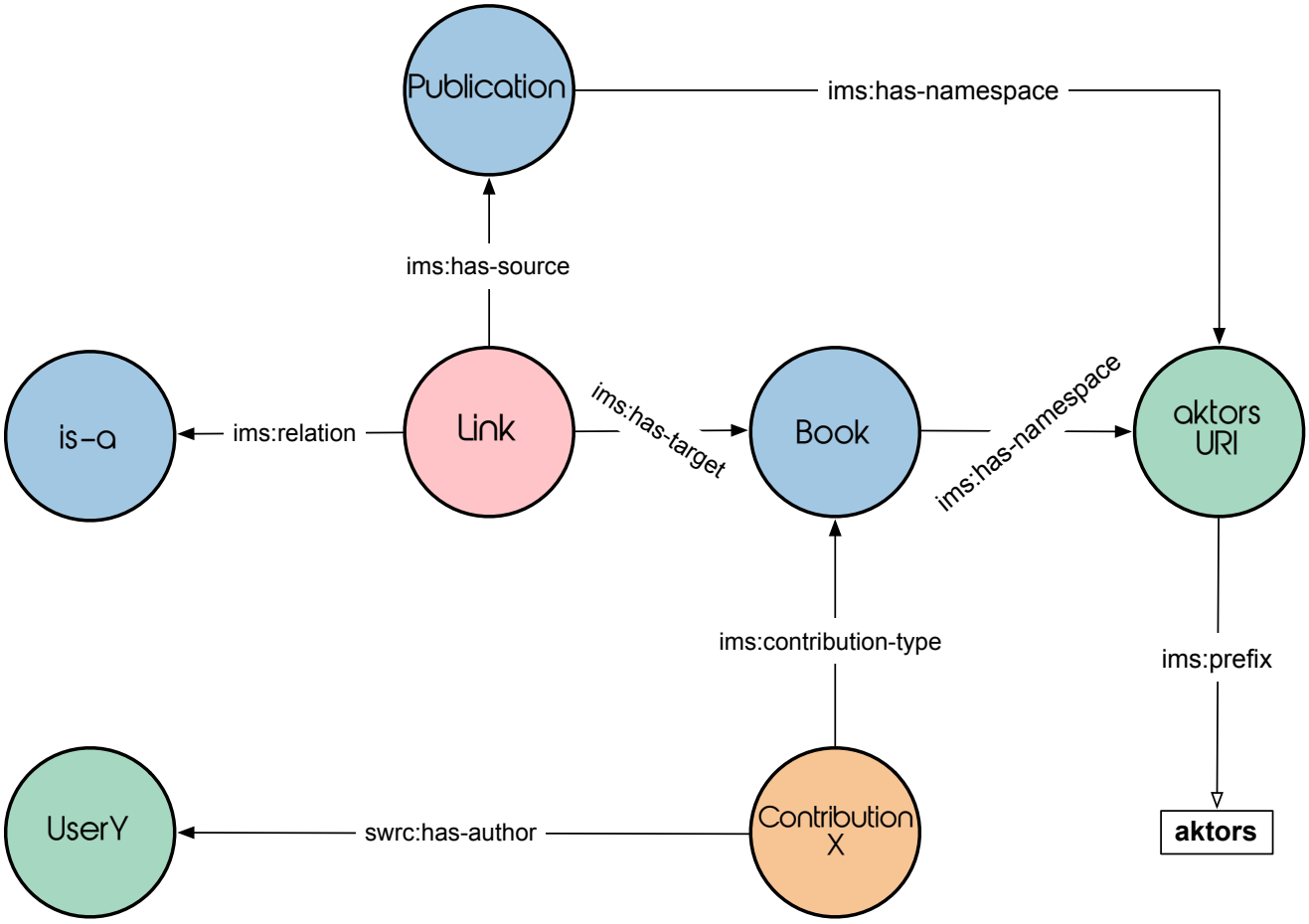
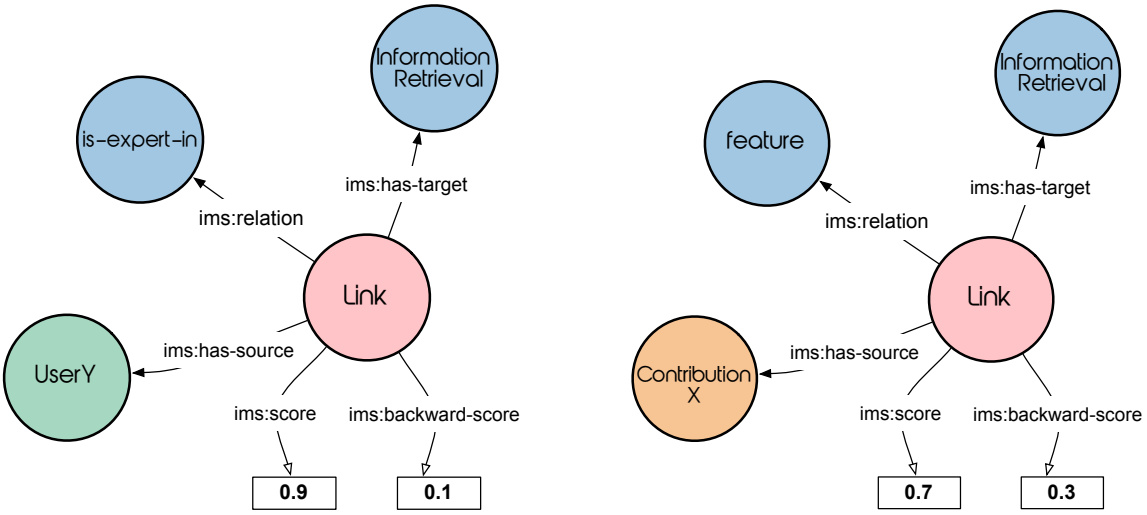


Fig. 6 The RDF graph of an instantiation of the model shown in Figure 5; furthermore it shows how the terms "Publication" and "Book" are associated with the terms in the Advanced Knowledge Technology reference ontology (i.e., aktors).

`contribution-type` property. Moreover, in the lower part of Figure 6 we can see how a contribution is associated with an author via the `swrc:has-author` property representing a user—i.e., "userY"—author of "contributionX".

`Link` has two datatype properties: `ims:score` and `ims:backward-score`, which allow us to add weights on any typed relationship; both score and backward score are `xsd:double` in the interval $[0, 1]$. Indeed, we can establish a relation between user and concept with two scores on it in order to say that a user is expert in a given scientific topic. This lets us define expert profiles; for instance, we can say that "userY is an expert in Information Retrieval" where "userY" is an instance of the `User` class and "information retrieval" is a term defined as an instance of `Concept`; the score represents the strength of the relation between a user and a concept, and the backward score represents the strength of the relation between a concept and a user. This means that the relationship between `User` and `Concept` is not symmetric; for instance, we can say that "UserY" is an expert in "Information Retrieval" with score 0.9 and this means that information retrieval is the main area of expertise for the user. On the other hand, there may be people more expert in information retrieval than "UserY", so the backward score can be set to be only 0.1, and this would mean that "UserY" is just one of the experts in "Information Retrieval" and that we expect to find out other users with a higher expertise level (backward score). The RDF graph of the user profile just described is shown in Figure 7(a).

In Figure 7(b) we can see another possible use of `Link`, in this case for representing the relationship between a contribution and a scientific topic. Indeed, semantic enrichment techniques are employed for extracting scientific topics from the data produced by the evaluation workflow and then relating them with pertinent contributions. We can see that "contributionX" is related to the scientific topic "Information Retrieval" via an `ims:relation` called "feature"; also in this case the typed relation between `contribution` and `concept` is weighted; the score is set to 0.7 meaning that "contributionX" mainly talks about "Information Retrieval", whereas the backward score is set to "0.3" meaning that



(a) An instantiation of the model shown in Figure 5 for representing an expert profile.

(b) An instantiation of the model shown in Figure 5 for associating a contribution with a scientific topic and vice versa.

Fig. 7 Two RDF graphs instantiating the model shown in Figure 5.

among contributions about "Information Retrieval", "contributionX" is not one of the most prominent ones.

5 Accessing the Experimental Data

The described RDF model has been realized by the DIRECT system which allows for accessing the experimental evaluation data enriched by the expert profiles created by means of the techniques that will be described in the next sections. This system is called LOD-DIRECT and it is available 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 expertise topics which are available as linked data as well.

At the time of writing, LOD-DIRECT allows access to 2,229 contributions, 2,334 author profiles and 2,120 expertise topics. Overall, 1,659 experts have been individuated and on average there are 8 experts per expertise topics (an expert can have more than one expertise of course).

LOD-DIRECT serializes and allows the access to the defined resources in several different formats such as XML, JSON, RDF+XML, Turtle[17] and Notation3 (n3)[18].

[17] `http://www.w3.org/TR/turtle/`
[18] `http://www.w3.org/TeamSubmission/n3/`

The URIs of the resources are constructed following the pattern:

`base-path/{resource-name}/{id};{ns}`

where,

- `base-path` is `http://lod-direct.dei.unipd.it`;
- `resource-name` is the name of the resource to be accessed as defined in the RDF model presented above;
- `id` is the identifier of the resource of interest;
- `ns` is the namespace of the resource of interest, this applies only for the namespace identifiable resources.

As an example, the URI corresponding to the contribution resource shown in Figure 2 with identifier `CLEF2012wn-RepLab-KarlgrenEt2012b` is:

`http://lod-direct.dei.unipd.it/contribution/CLEF2012wn-RepLab-KarlgrenEt2012b`

In Figure 8 we can see the Turtle serialization returned by the URI above. The serialization of a contribution is composed of four main parts: (i) the prefixes, reporting all the required information about the vocabularies adopted by the RDF model to represent the given resource; (ii) the authors of the contribution, which in this case are four comprising "Jussi Karlgren" who is the expert in "Reputation Management" reported in the use case in Figure 2; (iii) the serialization of the contribution itself, which includes information such as the title and the link to get the linked digital object; and, (iv) the metadata describing the RDF representation of the contribution.

prefixes

```
@prefix dc: <http://purl.org/dc/terms/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix aktors: <http://www.aktors.org/ontology/portal#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix ims: <http://ims.dei.unipd.it/data/rdf/> .
@prefix bibo: <http://purl.org/ontology/bibo/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix swrc: <http://swrc.ontoware.org/ontology#> .
```

authors

```
<http://lod-direct.dei.unipd.it/user/Fredrik+Olsson;http://
ims.dei.unipd.it/author/>
ims:file-metadata _:b0 ;
ims:has-namespace <http://lod-direct.dei.unipd.it/namespace/http://
ims.dei.unipd.it/author/> ;
ims:identifier "Fredrik+Olsson" .

<http://lod-direct.dei.unipd.it/user/Fredrik+Espinoza;http://
ims.dei.unipd.it/author/>
ims:file-metadata _:b0 ;
ims:has-namespace <http://lod-direct.dei.unipd.it/namespace/http://
ims.dei.unipd.it/author/> ;
ims:identifier "Fredrik+Espinoza" .

<http://lod-direct.dei.unipd.it/user/Magnus+Sahlgren;http://
ims.dei.unipd.it/author/>
ims:file-metadata _:b0 ;
ims:has-namespace <http://lod-direct.dei.unipd.it/namespace/http://
ims.dei.unipd.it/author/> ;
ims:identifier "Magnus+Sahlgren" .

<http://lod-direct.dei.unipd.it/user/Jussi+Karlgren;http://
ims.dei.unipd.it/author/>
ims:file-metadata _:b0 ;
ims:has-namespace <http://lod-direct.dei.unipd.it/namespace/http://
ims.dei.unipd.it/author/> ;
ims:identifier "Jussi+Karlgren" .

<http://lod-direct.dei.unipd.it/user/Ola+Hamfors;http://
ims.dei.unipd.it/author/>
ims:file-metadata _:b0 ;
ims:has-namespace <http://lod-direct.dei.unipd.it/namespace/http://
ims.dei.unipd.it/author/> ;
ims:identifier "Ola+Hamfors" .
```

contribution

```
<http://lod-direct.dei.unipd.it/contribution/CLEF2012wn-RepLab-
KarlgrenEt2012b>
ims:contribution-type <http://lod-direct.dei.unipd.it/concept/
Publication;http://www.aktors.org/ontology/portal%23> ;
ims:copyrighted "false" ;
ims:created "2013-05-19T17:01:05.644+02:00" ;
ims:file-metadata _:b0 ;
ims:last-modified "2013-05-19T17:01:05.644+02:00" ;
ims:link "http://www.clef-initiative.eu/documents/71612/155385/
CLEF2012wn-RepLab-KarlgrenEt2012b.pdf" ;
ims:owner <http://lod-direct.dei.unipd.it/user/root;http://
ims.dei.unipd.it/> ;
ims:title "Profiling Reputation of Corporate Entities in Semantic Space
" ;
swrc:has-author <http://lod-direct.dei.unipd.it/user/Magnus
+Sahlgren;http://ims.dei.unipd.it/author/> , <http://lod-
direct.dei.unipd.it/user/Jussi+Karlgren;http://ims.dei.unipd.it/author/
> , <http://lod-direct.dei.unipd.it/user/Fredrik+Espinoza;http://
ims.dei.unipd.it/author/> , <http://lod-direct.dei.unipd.it/user/
Fredrik+Olsson;http://ims.dei.unipd.it/author/> , <http://lod-
direct.dei.unipd.it/user/Ola+Hamfors;http://ims.dei.unipd.it/author/> .
```

metadata

```
<http://lod-direct.dei.unipd.it/namespace/http://
ims.dei.unipd.it/>
ims:file-metadata _:b0 ;
ims:identifier "http://ims.dei.unipd.it/" ;
ims:prefix "e6fe2c43" .

<http://lod-direct.dei.unipd.it/user/root;http://
ims.dei.unipd.it/>
ims:file-metadata _:b0 ;
ims:has-namespace <http://lod-direct.dei.unipd.it/
namespace/http://ims.dei.unipd.it/> ;
ims:identifier "root" .

<http://lod-direct.dei.unipd.it/namespace/http://
www.aktors.org/ontology/portal%23>
ims:file-metadata _:b0 ;
ims:identifier "http://www.aktors.org/ontology/portal%23" ;
ims:prefix "37675fe1" .


_:b0 dc:created "2015-11-06T15:55:20.052+01:00" ;
dc:creator "LOD DIRECT (Distributed Information Retrieval
Evaluation Campaign Tool) - Version 3.10" ;
dc:rights "Copyright (c) 2006-2015 - Information Management
Systems (IMS) Research Group (http://ims.dei.unipd.it/) -
Department of Information Engineering (http://
www.dei.unipd.it/) - University of Padua (http://
www.unipd.it/)" .

<http://lod-direct.dei.unipd.it/namespace/http://
ims.dei.unipd.it/author/>
ims:file-metadata _:b0 ;
ims:identifier "http://ims.dei.unipd.it/author/" ;
ims:prefix "9c5e2261" .

<http://lod-direct.dei.unipd.it/concept/Publication;http://
www.aktors.org/ontology/portal%23>
ims:file-metadata _:b0 ;
ims:has-namespace <http://lod-direct.dei.unipd.it/
namespace/http://www.aktors.org/ontology/portal%23> ;
ims:identifier "Publication" .
```

Fig. 8 The Turtle serialization of a contribution (i.e. `CLEF2012wn-RepLab-KarlgrenEt2012b`) returned by the LOD-DIRECT system.

The metadata reported in Figure 8 is an instance of the metadata returned for each resource in the LOD-DIRECT system; this metadata is "intended as a bridge between the publishers and users of RDF data" as in the case of VoID (Vocabulary of Interlinked Datasets)[19]. As a matter of fact, the LOD-DIRECT system employs the VoID description principles; for instance, the author and the rights related to the considered resource are described by means of the Dublin Core vocabulary (i.e. `dc:creator` and `dc:rights`) as prescribed by the VoID specification.

LOD-DIRECT comes with a fine-grained access control infrastructure which takes care of monitoring the access to the various resources and functionalities offered by the system. On the basis of the requested operation, it performs:

- authentication, i.e., it asks for user credentials before allowing to perform an operation;
- authorization, i.e., it verifies that the user currently logged in holds sufficient rights to perform the requested operation.

The access control policies can be dynamically configured and changed over the time by defining roles, i.e., groups of users, entitled to perform given operations.

[19] `http://www.w3.org/TR/void/`

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 fine-grained access control to resources is managed via groups of users, which can have different access permissions. The general rules are as follows:

- private resources: they can be read and modified only by the owner of the resource;
- shared resources: they can be read and modified by the owner of the resource; then, a list of groups can share the resource with different access permissions, namely "read only", which means that the users of that group can only read but not modify the resource, and "read/write", which means that the users of that group can read and modify the resource;
- public resources: they can be read by everybody; they can be read and modified by the owner of the resource; then, a list of groups can share the resource with different access permissions, namely "read only", which means that the users of that group can only read but not modify the resource, and "read/write", which means that the users of that group can read and modify the resource.

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.

6 Semantic Enrichment

In this section we describe several methods for semantically enriching experimental IR data modelled as described above, by analysing unstructured data available in scientific publications. Figure 9 presents an overview of the semantic enrichment of documents and authors based on term and topical hierarchy extraction. First, we propose a method to automatically extract expertise topics from a domain-specific collection of publications using an approach for term extraction in Section 6.1. Then, we present a preliminary approach for enriching expertise topics by grounding them in the LOD cloud in Section 6.2. An approach for expert profiling based on automatically extracted expertise topics is discussed in Section 6.3. In Section 6.4 we present several measures that can be used to rank experts for a given topic making use of an automatically extracted hierarchy of terms.

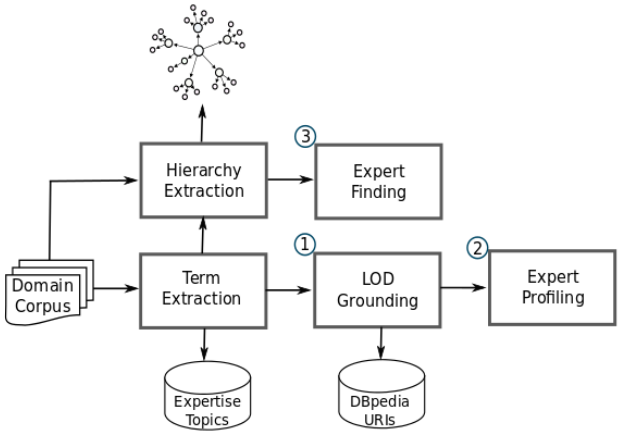


Fig. 9 Data flow of the semantic enrichment approach

6.1 Expertise topic extraction

Topic-centric approaches for expert search emphasize the extraction of keyphrases that can succinctly describe expertise areas, also called expertise topics, using term extraction techniques (Bordea et al., 2012). An advantage of a topic-centric approach compared to previous work on expert finding (Balog et al., 2012) is that topical profiles can be constructed directly from text, without the need for controlled vocabularies or for manually identifying terms. Expertise topics are extracted from a domain-specific corpus using the following approach. First, candidate expertise topics are discovered from text using a syntactic description for terms (i.e., nouns or noun phrases) and contextual patterns that ensure that the candidates are coherent within the domain. A domain model is constructed using the method proposed in Bordea et al. (2013b) and then noun phrases that include words from the domain model or that appear in their immediate context are selected as candidates. Candidate terms are further ranked using the scoring function s, defined as:

$$s(\tau) = |\tau| \log f(\tau) + \alpha e_\tau \quad (1)$$

where τ is the candidate string, $|\tau|$ is the number of words contained by candidate τ, f is its frequency in the corpus, and e_τ is the number of terms that embed the candidate string τ. The parameter α is used to linearly combine the embeddedness score e_τ. In Table 4 we report the top ranked expertise topics extracted from IR publications by using the described method. These topics describe core concepts of the domain such as *search engine*, *IR system*, and *retrieval task*, as well as prominent subfields of the domain including *image retrieval*, *machine translation*, and *question answering*.

Only the best 20 expertise topics are stored for each document, ranking expertise topics based on their over-

all score $s(\tau)$ multiplied with their *tf-idf* score. In this way, each document is enriched with keyphrases, taking into consideration the quality of a term for the whole corpus in combination with its relevance for a particular document.

Table 4 Top 20 expertise topics extracted from IR scientific publications

Rank	Expertise Topic
1	information retrieval
2	image retrieval
3	retrieval systems
4	search engine
5	information retrieval system
6	retrieval task
7	QA system
8	query expansion
9	language model
10	text retrieval
11	target language
12	training data
13	retrieval model
14	visual features
15	question answering system
16	Natural Language
17	machine translation
18	relevance feedback
19	IR system
20	annotation task

6.2 Enriching expertise topics using LOD

Expertise topics can be used to provide links between IR experimental data and other data sources. These links play an important role in cross-ontology question answering, large-scale inference and data integration (Ngonga Ngomo, 2012). Also, existing work on using knowledge bases in combination with IR techniques for semantic query expansion shows that background knowledge is a valuable resource for expert search (Demartini, 2007; Thiagarajan et al., 2008). Additional background knowledge, as found on the LOD cloud [20], can inform expert search at different stages. For example, manually curated concepts can be leveraged from a large number of domain-specific ontologies and thesauri. Also, the LOD cloud contains a large number of datasets about scientific publications and patent descriptions that can be used as additional evidence of expertise.

A first step in the direction of exploiting this potential is to provide an entry point in the LOD cloud through DBpedia [21], one of the most widely connected datasources, that is often used as an entry point in the LOD cloud. Two promising approaches for semantic term grounding on DBpedia are described and evaluated in Section 7.2.1. Our goal is to associate as many terms as possible with a concept from the LOD cloud through DBpedia URIs—as shown in the use-case presented in Section 3. Where available, concept descriptions are collected as well and used in our system. Initially we find all candidate URIs using the following DBpedia URI pattern.

http://dbpedia.org/resource/{*DBpedia_label*}

Here *DBpedia_concept_label* is the expertise topic as extracted from our corpus. A large number of candidates are generated starting from a multi-word term as each word from the concept label can start with a letter in lower case or upper case in the DBpedia URI. As an example, let us consider the expertise topic *"Natural Language Processing"*, all possible case variations are generated to obtain the following URI:

http://dbpedia.org/page/Natural_language_processing

To ensure that only DBpedia articles that describe an entity are associated with an expertise topic, we discard category articles and we consider only articles that match the `dbpedia-owl:title` or the final part of the candidate URI with the topic. Multiple morphological variations are extracted and stored from our corpus for each expertise topic. Each of these variations is used to search for a URI, increasing the number of matches.

6.3 Expert profiling

Expertise profiles are brief descriptions of a person's expertise and interests, that can inform the selection of experts in different scenarios. Whenever we refer to an expertise profile throughout this work, we mean a topical profile. Although a person frequently writes about a subject area, the way they combine this area with other topics is more interesting, because a person is rarely an expert on every aspect of a topic (Mimno and McCallum, 2007). In Berendsen et al. (2013), several requirements are identified for expertise profiles including coherence, completeness, conciseness, and diversity. The same study states that an important requirement for expertise topics is that they have to be at the right level of specificity.

Following Balog and de Rijke (2007), we define a topical profile of a candidate as a vector of expertise topics along with scores that measure the expertise of a candidate. Therefore, the expertise profile p of a researcher r is defined as:

$$p(r) = \{s(r, t_1), s(r, t_2), ..., s(r, t_n)\} \quad (2)$$

[20] Linked Data: `http://linkeddata.org`
[21] DBpedia: `http://dbpedia.org/`

where t_1, t_2,...,t_n are the expertise topics extracted from a domain specific corpus.

A first step in constructing expertise profiles is to identify terms that are appropriate descriptors of expertise. A large number of expertise topics can be extracted for each document, but only the top-ranked keyphrases are considered for expert profiling, as described in the previous section. Once a list of expertise topics is identified, we proceed with assigning scores to each expertise topic for a given expert. We rely on the notion of relevance, effectively used for document retrieval, to associate expertise topics with researchers. Researchers' interests and expertise are inferred based on their scientific contributions. Each expertise topic mentioned in one of these contributions is assigned to their expertise profile using an adaptation of the standard IR measure tf-idf (Baeza-Yates and Ribeiro-Neto, 1999). The set of contributions authored by a researcher is aggregated into a virtual document, that allows us to compute the relevance of an expertise topic for each researcher. In the case of multi-author publications, the authors are considered to contribute equally to the each of the topics mentioned in the paper. This is not always the case, therefore profiles tend to be more accurate when multiple publications authored by a person are available. An expertise topic is added in the expertise profile of a researcher using the following scoring function:

$$s_{ep}(r,t) = termhood(t) \cdot tfirf(t,r) \quad (3)$$

Where $s_{ep}(r,t)$ represents the score for an expertise topic t and a researcher r, $termhood(t)$ represents the score computed in Equation 1 for the topic t and $tfirf(t,r)$ stands for the tf-idf measure for the topic t on the aggregated document of researcher r. In this way, we construct profiles with terms that are representative for the domain as well as highly relevant for a given researcher.

6.4 Expert finding

Expert finding is the task of identifying the most knowledgeable person for a given expertise topic. In this task, several competent people have to be ranked based on their relative expertise on a given expertise topic. Documents written by a person can be used as an indirect evidence of expertise, assuming that an expert often mentions his areas of interest. We rely on the tf-irf measure described in the previous section to measure the relevance of a given expertise topic for a researcher. Each researcher is represented by an aggregated document that is constructed by concatenating all the documents authored by that person. Therefore, the relevance score $R(r,t)$ that measures the interest of a researcher r for a given topic t is defined as:

$$R(r,t) = tfirf(t,r) \quad (4)$$

Expertise is closely related to the notion of experience. The assumption is that the more a person works on a topic, the more knowledgeable they are. We estimate the experience of a researcher on a given topic by counting the number of publications that have the topic assigned as a top ranked keyphrase. Let $D_{r,t}$ be the set of documents authored by researcher r, that have the expertise topic t as a keyphrase. Then, the experience score $E(r,t)$ is defined as:

$$E(r,t) = |D_{r,t}| \quad (5)$$

Where $|D_{r,t}|$ is the cardinality, or the total number of documents, in the set of documents $D_{r,t}$. It can be argued that it is not only the number of publications that indicates expertise, but the quality of those publications as well. We leave for future work the integration of publication impact in this score, measured using citation counts modeled by the DIRECT conceptual model and available in the RDF graph of the exposed experimental data.

Relevance and expertise measure different aspects of expertise and can be combined to take advantage of both features as follows:

$$RE(r,t) = R(r,t) \cdot E(r,t) \quad (6)$$

Both the relevance score and the experience score rely on query occurrences alone. But a topical hierarchy can provide valuable information about hierarchical relations between expertise topics, and can improve expert finding results. Taxonomies are not always available and are difficult to maintain, therefore we consider an automatic approach for extracting hierarchical relations. Take for example the topical hierarchy presented in Figure 10, which was automatically constructed using publications from the CLEF evaluation campaign using the method proposed in Hooper et al. (2012). When searching for experts in *image retrieval*, we can make use of the information that *image annotation* and *visual features* are closely related expertise topics that are subordinated to the topic of interest. In the same way, when searching experts on the expertise topic *question answering* we can use information about the subordinated terms *QA system* and *answer extraction*.

In the case that the subtopics of an expertise topic are known, we can evaluate the expertise of a person based on their knowledge of the more specialised fields.

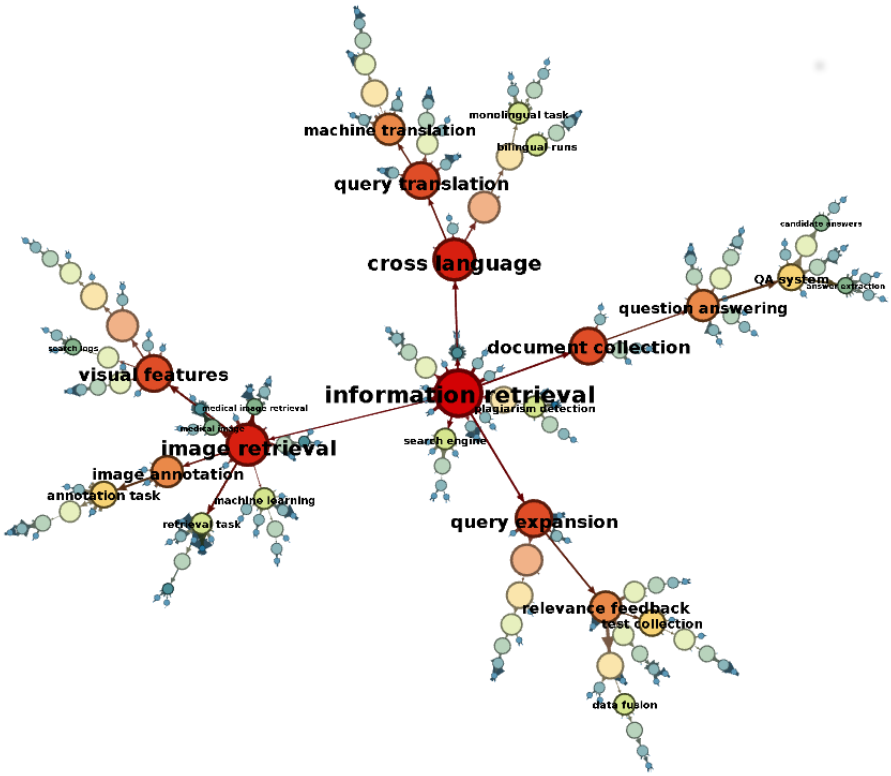


Fig. 10 Topical hierarchy automatically constructed for the CLEF evaluation campaign

A previous study showed that experts have increased knowledge at more specific category levels than novices (Tanaka and Taylor, 1991). We introduce a novel measure for expertise called *Area Coverage* that measures whether an expert has in depth knowledge of an expertise topic, using an automatically constructed topical hierarchy. Let $Desc(t)$ be the set of descendants of a node t in the topical hierarchy, then the Area Coverage score $C(i,t)$ is defined as:

$$C(i,t) = \frac{|\{t' \in Desc(t) : t \in p(i)\}|}{|Desc(t)|} \quad (7)$$

where $p(i)$ is the profile of an individual i constructed using the method presented in the previous section. In other words, Area Coverage is defined as the proportion of a term's descendants that appear in the profile of a person. For expertise topics that have no descendent, Area Coverage is defined as 1.

Finally, the score $REC(i,t)$ used to rank people for expert finding is defined as follows:

$$REC(i,t) = RE(i,t) \cdot C(i,t) \quad (8)$$

This score combines several performance indicators, measuring the expertise of a person based on the relevance of an expertise topic, the number of documents about the given topic, as well as his depth of knowledge of the field, called Area Coverage.

7 Experimental evaluation

7.1 Experimental setup

7.1.1 Expert search datasets

Evaluating expert search systems remains a challenge, despite a number of data sets that have been made publicly available in recent years (Bailey et al., 2007; Balog et al., 2007; Soboroff et al., 2007). Traditionally, relevance assessments for expert finding were gathered either through self-assessment or based on opinions of co-workers. On the one hand, self-assessed expert profiles are subjective and incomplete, while on the other hand opinions of colleagues are biased towards their social and geographical network. We address these limitations by exploiting expertise data generated in a peer-review setting (Bordea et al., 2013a). Our aim is to collect a representative dataset of experts in Information Retrieval along with their publications and expertise topics. We consider conference workshops in the related fields of IR, DL, and *Recommender Systems (RS)*. About 25 thousand publications are gathered along with data about 60 workshops. Each workshop is associated on average with 15 experts and almost 500 expertise topics are manually extracted to describe these events.

To construct a test collection covering all these research fields, we used the DBLP Computer Science Bibliography[22]. Our initial motivations for constructing a test collection around DBLP are two-fold: (1) the fields of IR, DL, and RS are well-covered in DBLP, and (2) a special version of the DBLP data set, augmented with citation information, is available from the team behind ArnetMiner, which allows for investigations into the use of citation information for expert search.

To make the augmented DBLP collection suited to expert search evaluation, we need realistic topic descriptions as relevance judgments at the expert level. Workshops organized at major conferences covering the fields of IR, DL, and RS are used to collect relevance judgments. To identify relevant workshops, we visited the websites of the CIKM, ECDL, ECIR, IIiX, JCDL, RecSys, SIGIR, TPDL, WSDM, and WWW conferences, which have substantial portions of their program dedicated to IR, DL, and RS. We collect links to workshop websites for all workshops organized at those conferences between 2001 and 2012. This resulted in a list of 60 different workshops with websites.

As a starting point, a test collection covering the aforementioned fields is constructed by using the augmented DBLP data set released by the team behind ArnetMiner. This data set is a October 2010 crawl of of the DBLP data set containing 1,632,442 different papers with 2,327,450 citation relationships between papers in the data set[23]. As this augmented data set contains publications from all fields of computer science, we filtered out all publications not belonging to IR, DL, and RS by restricting the collection to publications in relevant journals, conferences, and workshops. This step and all of the steps listed below were completed in June 2012.

The list of relevant venues was created in two steps. First, we generated a list of *core venues* by extracting all papers published at conferences used for topic creation: CIKM, ECDL, ECIR, IIiX, JCDL, RecSys, SIGIR, TPDL, WSDM, and WWW. We select these conferences, because as hosts to the topic workshops, they are likely to be relevant venues for PC members to publish in. This resulted in a data set containing 9,046 different publications from these core venues. However, restricting ourselves to these venues alone means we could be missing out on experts that tend to publish more in journals and workshops. We therefore extend the list of core venues with other venues tracked by DBLP that also have substantial portions of their program dedicated to IR, DL, and RS. Venues that only feature incidental overlap with IR, such as the Semantic Web conference, are not included. We also exclude venues that did not have 5 publications or more in the augmented DBLP data set. While this does exclude the occasional on-topic publication in venues that are pre-dominantly about other topics, we believe that this strategy will cover the majority of relevant publications. This additional filtering step results in a final list of 78 *curated venues* (core plus additional) covering a total of 24,690 publications.

In addition to citation information, the augmented DBLP data set was also extended with abstracts wherever available. However, the team behind ArnetMiner was only able to add abstracts for 33.7% of the 1.6 million publications (and 43.5% of the 24,690 publications in our test collection). We therefore attempted to download the full-text versions of all 24,690 publications using Google Scholar. We constructed a search query consisting of the last name of the first author and the full title without surrounding quotes[24]. We then extracted the download link from the top result returned by Google Scholar (if available). We were able to find download URLs for 14,823 of the 24,690 publications in our filtered DBLP data set for which a recall of 60.04%, where recall is defined as the percentage of papers in our filtered DBLP data set for which we could find download URLs. While this is not as high as we would like, it does represent a substantial improvement over the percentage of abstracts present in the augmented DBLP data set. Moreover, a recall rate of 100% is impossible to achieve as tutorials, keynote abstracts, and even entire proceedings are typically not available online in full-text, but they are present in the DBLP data set.

Around 90.15% of the download URLs obtained in this manner were functional, which means we were able to download full-text publication files for 13,363 publications (or 54.12% of our entire curated data set). We performed a check of 100 randomly selected full-text files to see if these are indeed the publications we are looking for and achieved a precision of 97% on this sample. We therefore assume that the false positive rate of our approach is acceptably low. This augmented DBLP collection is publicly available[25].

Beside the Information Retrieval dataset described above, we report results obtained for similar datasets in two other Computer Science fields, including Semantic Web and Computational Linguistics. Table 5 gives an overview of the considered datasets in terms of number of documents, workshops, authors and expertise topics.

[22] `http://dblp.uni-trier.de/`

[23] Available at `http://arnetminer.org/DBLP_Citation`, last accessed July 9, 2013.

[24] A preliminary test on just the publications from the core venues showed that adding quotes around the publication title decreased recall from 80.3% to 70.86%.

[25] `http://toinebogers.com/?page_id=660`

	IR	CL	SW
#documents	24,690	10,921	2,311
% of full-text documents	54.1%	100%	55%
#workshops	60	340	190
#unique authors	26,098	9,983	4,480
#authors/document	2.7	2.2	3.3
#experts/workshop	14.9	25.8	24.9
#expertise topics	488	4,660	6,751

Table 5 Overview of workshop based test collections for Information Retrieval (IR), Computational Linguistics (CL), and Semantic Web (SW)

These domain-specific datasets contain a large amount of scientific publications that are focused on a given field of research. This allows us to investigate expertise in a given research community, while previous studies on Expert Search put more effort into analysing expertise inside knowledge-intensive organisations. The UvT dataset, introduced in Balog et al. (2007), contains information about the employees of Tilburg University, that was collected from a publicly accessible expertise database. The UvT dataset is more heterogeneous than the workshop datasets, as it gathers information from manually provided summaries of research and courses, personal homepages, as well as publications. Table 6 gives an overview of the size of the UvT dataset. The UvT dataset is topically more diverse than the datasets presented in the previous section, covering broad areas of study such as economics, law, information technology, public administration or criminology. Although expertise topics are available in Dutch and English, in our experiments we considered only 981 expertise topics available in English.

	RD	CD	PUB	HP
#documents	316	840	27,682	6.724
#people	316	318	734	318

Table 6 Overview of the UvT Expert Dataset, including Research Descriptions (RD), Course Descriptions (CD), Publications (PUB), and Personal Homepages (HP)

About 7% of the publications are available as full content, with most publications being available as citations only. The large and diverse number of expertise topics combined with the limited availability of textual descriptions leads to challenges related to data sparseness. Nevertheless, the expert finding and expert profiling tasks are easier on the UvT dataset. This is due to the fact that most documents are high quality summaries of expertise and that there are a relatively smaller number of people in the dataset. Additionally, there is a small number of overlapping expert profiles, because in a university less people have similar interests than in a research community.

7.1.2 Baseline approaches

The approaches proposed in this section are evaluated against two IR methods for expert finding and expert profiling (Balog et al., 2009). Both methods model documents and expertise topics as bags of words and take a generative probabilistic approach, ranking expertise topics t by the probability $P(t|i)$ that they are generated by the individual i (Balog et al., 2009). The same probability is used for ranking expertise topics in a person's profile, as well as for finding knowledgeable people for expert finding. The first model constructs a multinomial language model θ_i for each individual, over the vocabulary of documents authored by them. This is similar to our approach that computes the relevance of a topic for an individual on a document that aggregates all the documents authored by that person.

The assumption is that expertise topics are sampled independently from this multinomial distribution. Therefore, the probability $P(t|i)$ can be computed as:

$$P(t|i) = P(t|\theta_i) = \prod_{w \in t} P(w|\theta_i)^{n(w,t)} \tag{9}$$

where $n(w,t)$ is the number of times the word w appears in the expertise topic t. Smoothing using collection word probabilities is applied to estimate $P(w|\theta_i)$. The smoothing parameters are estimated with an unsupervised method, using Dirichlet smoothing and the average number of words associated with people as the smoothing parameter.

The second model considered as baseline, that is also introduced in (Balog et al., 2009), estimates a language model θ_d for each document from the set D_i of documents authored by the individual i. Words from an expertise topic t are sampled independently, summing the probabilities to generate an expertise topic for each of these documents. In this case, the probability $P(t|i)$ is calculated using the following equation:

$$P(t|i) = \sum_{d \in D_i} P(t|\theta_d) = \sum_{d \in D_i} \prod_{w \in t} P(w|\theta_d)^{n(w,t)} \tag{10}$$

Again, the probability $P(w|\theta_d)$ is estimated by using the same unsupervised smoothing method. In this case, the smoothing parameter for Dirichlet smoothing is the average document length in the corpus.

7.1.3 Evaluation measures

Given the tasks at hand, several evaluation measures for document retrieval can be used. The expert profiling and the expert finding tasks are evaluated based on the quality of ranked lists of expertise topics and of experts, respectively. From an evaluation point of view, this is not different from evaluating a ranked list of documents with binary relevance judgments—i.e., a document is either relevant or not with respect to a given topic. The most basic evaluation measures used in IR are precision and recall. The first measure is given by the ratio between the number of relevant documents retrieved and the total number of retrieved documents. The second is given by the ratio between the number of relevant documents retrieved and the total number of relevant documents for a given topic. Other frequently used effectiveness measures include:

Precision at N (P@N) (van Rijsbergen, 1979) This is the precision computed when N results are retrieved, which is usually used to report early precision at top 5, 10, or 20 results.

Average Precision (AP) (Harman, 2011) Precision is calculated for every retrieved relevant result and then averaged across all the results.

Reciprocal Rank (RR) (Chapelle et al., 2011) This is the reciprocal of the first retrieved relevant document, which is defined as 0 when the output does not contain any relevant documents.

To get a more stable measurement of performance, these measures are commonly averaged over the number of queries. In our experiments, we report the values for the Mean Average Precision (MAP), and Mean Reciprocal Rank (MRR). In this setting, recall is less important than achieving a high precision for the top ranked results, because it is more important to recommend true experts than to find all experts in a field.

7.2 Experiments

7.2.1 Semantic grounding of expertise topics

Two approaches for grounding expertise topics on DBpedia are evaluated in this section. The first approach matches a candidate DBpedia URI with an expertise topic, using the string as it appears in the corpus. The second approach makes use of the lemmatised form of the expertise topic. Stemming was also considered but this approach resulted in a decrease in performance, as stems are more ambiguous [26]. In order to evaluate our URI discovery approach, we build a small gold standard dataset by manually annotating 186 expertise topics with DBpedia URIs. First of all, we note that about half of the analysed expertise topics have a corresponding concept in DBpedia. One of the main reasons for the low coverage is that DBpedia is a general knowledge datasource that has a limited coverage of specialised technical domains.

Table 7 Precision and recall for DBpedia URI extraction

Approach	Precision	Recall	F-score
String Matching	0.96	**0.93**	0.94
Lemmatisation	**0.99**	0.90	0.94

Although both approaches achieve similar results in terms of F-score, the approach that makes use of lemmatisation (A2) achieves better precision, as can be seen in Table 7. Surprisingly, using lemmatization achieves a lower recall but higher precision but this might be due to the small size of the dataset. To extract descriptions or definitions of concepts we rely on the `dbpedia-owl:abstract` property, or the `rdfs:comment` property in the absence of the former. For now we are interested in English definitions, therefore we consider triples tagged with the property `lang='en'` alone. Even though English descriptions are available for a larger number of topics, this tag is not always present. Therefore, we can only retrieve descriptions for a smaller number of topics. A manual analysis of matching errors showed that expertise topics that include an acronym (e.g. "NLG system" instead of "Natural Language Generation system") are more difficult to associate with a DBpedia concept, as often acronyms are ambiguous.

Other general purpose data sources, such as Freebase [27], or domain-specific data sources can be linked in a similar manner. A complex problem that we do not address in this work is the disambiguation of an expertise topic when multiple concepts from different domains can be matched. Usually, DBpedia provides a disambiguation page for such cases. In our implementation we did not analyse concepts that redirect to a disambiguation page, grounding only those expertise topics that are specific enough to be used in a single domain.

[26] An approach based on a semantic web search engine that uses keyphrase search to find structured data was also considered, restricting the search to the DBPedia domain. The results were disappointing, because only a limited number of retrieved results can be analysed. Often, the relevant DBpedia concept does not appear in the top results.

[27] `http://www.freebase.com/`

7.2.2 Expert profiling

The topic-centric approach (TC) for expert profiling proposed in Section 6.3 can be applied for expert profiling without the need for controlled vocabularies, as expertise topics are directly extracted from text. Instead, the language modelling approach used as a baseline in this section, can only be used on datasets where such resources are readily available. The results for the expert profiling task on the IR dataset are presented in Table 8.

Dataset	Measure	LM1	LM2	TC
	MAP	0.0256	0.0233	**0.0392**
CL	MRR	0.1857	0.2044	**0.2767**
	MAP	0.0082	0.0088	**0.0369**
SW	MRR	0.1271	0.1161	**0.3437**
	MAP	0.1052	**0.1679**	0.0879
IR	MRR	**0.3761**	0.3677	0.3364
	MAP	0.1299	**0.1380**	0.0459
UvT	MRR	0.3066	**0.3136**	0.1662

Table 8 Expert profiling results for the language modelling approach (LM) and the topic centric approach (TC)

The language modelling approaches achieve better results on the IR and the UvT datasets, with the LM2 approach outperforming the LM1 approach on most measures. The gap between the language modelling approaches and the TC approach is more narrow on the IR dataset. Not surprisingly, our method for extracting expertise topics is under-performing when applied to a corpus that covers diverse expertise areas, such as the UvT dataset. Another difference between these datasets is the number of documents that are available for each person. The LM1 and LM2 models achieve the worse results on the SW dataset, where only 8% of the people are associated with more than 3 documents.

7.2.3 Expert finding

We compare several topic-centric methods for expert finding with two language-modelling baselines. The results for the expert finding task are presented in Table 9. The expert finding methods evaluated in this section include Experience (E), Relevance and Experience (RE) and Relevance, Experience and Area Coverage (REC). These methods are described by Equations 5, 6, and 8 respectively, in Section 6.4. The Area Coverage measure makes use of a topical hierarchy. Therefore we automatically construct a topical hierarchy for IR using the method proposed in Hooper et al. (2012).

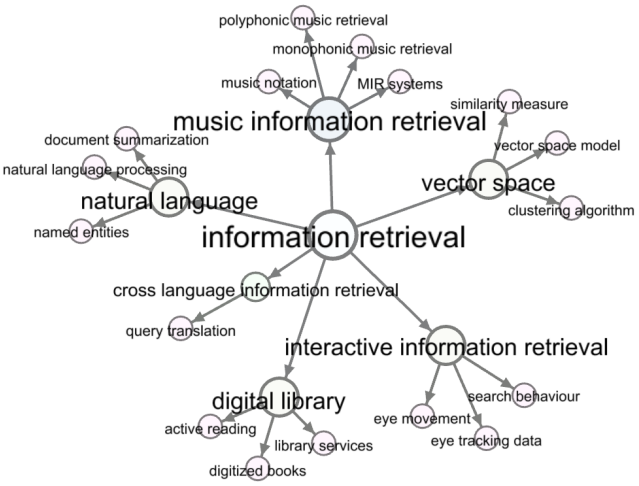


Fig. 11 Sample hierarchical relations for the IR domain

Figure 11 shows a small extract from this hierarchy that correctly identifies "information retrieval" as the root of the taxonomy as well as several subfields including "digital libraries", "interactive information retrieval", and "cross language information retrieval".

A short summary of the constructed topical hierarchies for each domain is presented in Table 10. Depending on the number of documents available in each dataset, a different number of expertise topics is extracted and subsequently considered for constructing a topical hierarchy. The CL dataset is the largest dataset, allowing us to filter edges in a pre-processing step based on the number of documents that provide evidence for the relation. An edge is added in the noisy graph only if at least three different documents provide evidence for the relation. This setting is not used for smaller datasets because it reduces the number of edges and the connectivity of the graph. For the same reason, the window size used to count co-occurrences of terms is larger for the smaller datasets than for the CL dataset. The topical hierarchy constructed for the IR domain is constructed by considering all the co-occurrences between two expertise topics in a window of 5 words. A larger window size would increase the number of edges but the relations would be less reliable. Considering more than 98% of the nodes are connected by an edge, we did not consider increasing the window size. Figure 12 presents an overview of node degree in the Information Retrieval hierarchy. More than half of the terms are specific terms that have no descendants, but a considerable number of nodes have several child nodes.

We note that topic-centric approaches (E, RE, REC) outperform language modelling approaches on domain-specific datasets such as the CL, SW, and IR datasets. Our experimental results lead us to the conclusion that

Dataset	Measure	LM1	LM2	E	RE	REC
CL	MAP	0.0071	0.0056	0.0335	0.0335	**0.0340**
	MRR	0.0631	0.0562	0.2734	0.2738	**0.2754**
	P@5	0.0202	0.0173	0.1340	0.1339	**0.1347**
SW	MAP	0.0070	0.0067	**0.0327**	0.0305	0.0314
	MRR	0.0528	0.0522	**0.2262**	0.2115	0.2095
	P@5	0.0182	0.0188	**0.1065**	0.0967	0.0994
IR	MAP	0.0599	0.0402	0.1592	**0.1669**	0.1657
	MRR	0.1454	0.1231	0.4056	**0.4141**	0.4120
	P@5	0.0614	0.0485	0.1771	0.1771	**0.1783**
UvT	MAP	**0.2009**	0.1994	0.1155	0.1151	0.1158
	MRR	0.3551	**0.3571**	0.2298	0.2266	0.2281
	P@5	**0.1357**	0.1347	0.0850	0.0846	0.0841

Table 9 Expert finding results for the language modelling approach (LM), Experience (E), Relevance and Experience (RE), and Relevance, Experience and Area Coverage (REC)

Dataset	CL	SW	UvT	IR
#Nodes	15,000	5.000	5,000	4,000
#Edges	14,976	4,506	4.939	3,939
#Min Docs	3	1	1	3
Window size	5	50	50	5

Table 10 Graph size for topical hierarchies constructed for Computational Linguistics (CL), Semantic Web(SW), Information Retrieval (IR), and Tilburg University (UvT)

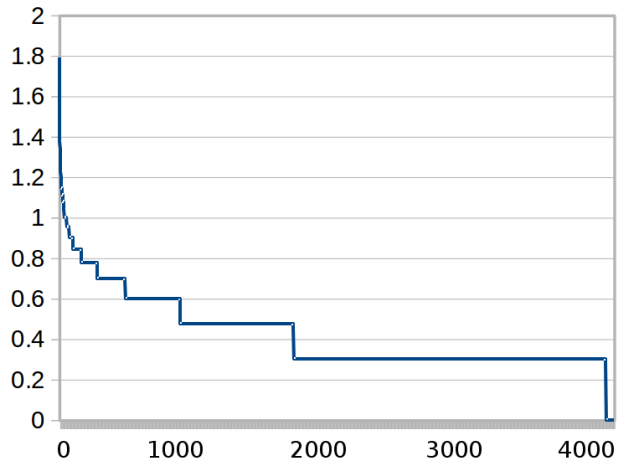


Fig. 12 Overview of node degree for the Information Retrieval hierarchy (logarithmic scale)

the more specialised a dataset is, the less reliable relevance-based assessment of expertise is. In the case of the Semantic Web dataset, which is the most focused dataset, using the relevance-based measure (RE) even decreases performance compared to the expertise score (E). Language modelling approaches outperform topic-centric approaches on the UvT dataset alone, which is the most broad dataset among the four considered datasets. This is because expertise profiles have a larger degree of overlap when dealing with focused datasets that describe a narrow domain. For example, it is easier to distinguish between experts in history and mathematics using relevance-based methods, but more difficult to distinguish between two experts in Semantic Web that address similar topics in their publications.

Using a topical hierarchy by computing Area Coverage improves the results across all datasets except the IR dataset, in terms of MAP. In terms of P@5, the results are improved on all datasets except on the UvT dataset. These results confirm our hypothesis that automatically constructed topical hierarchies can inform expert finding.

8 Conclusion

In this paper we discussed the data modelling and the semantic enrichment of IR experimental data, as produced by large-scale evaluation campaigns. We described in detail the evaluation workflow used for information access systems and we proposed an RDF model for two areas of the workflow, namely resource management and scientific production. This model is used as a common basis for semantic enrichment and for augmenting the discoverability, accessibility and re-usability of the experimental data. Unstructured data in the form of scientific publications were used to inform the extraction of various types of semantic enrichment. Expertise topics were automatically extracted and used to describe documents and to create expert profiles. Several topic-centric measures for expert finding were proposed, allowing users to identify knowledgeable members of the community. In this way we created new relationships among existing data, allowing a more meaningful interaction with them.

We introduced an evaluation dataset for expert search in IR, relying on scientific publications available online and on implicit expertise information about workshop committee members. Our experiments show that it is possible to construct expertise profiles using automatically extracted expertise topics and that topic-centric approaches for expert finding outperform state of the art language modelling approaches on most of the considered datasets.

In particular, besides the methodological contributions described above, the main re-usable deliverables of the paper are:

- an accurate RDF data model for describing IR experimental data in detail, available at `http://ims.dei.unipd.it/data/rdf/direct.3.10.ttl`;
- a dataset about CLEF contributions, extracted expertise topics and related expert profiles, developed according to the methods proposed in the paper;
- the online accessible LOD DIRECT system, available at `http://lod-direct.dei.unipd.it/`, to access the above data in different serialization formats, RDF+XML, Turtle, N3, XML and JSON.

Future work will concern the application of these semantic modeling and automatic enrichment techniques to other areas of the evaluation workflow. For example, expert profiling and topic extraction could be used to automatically improve and enhance the descriptions of the single experiments submitted to an evaluation campaign, which are typically not very rich and often cryptic—for example "second iteration with tuned parameters" as description—and to automatically link experiments to external resources, e.g., describing the used components, such as stemmers or stop lists, and systems. Finally, the RDF model defined within DIRECT opens up the possibility of integrating established DL methodologies for data access and management which increasingly exploit the LOD paradigm (Di Buccio et al., 2013; Hennicke et al., 2011; Lagoze et al., 2008; Stasinopoulou et al., 2007). This would enable broadening the scope and the connections between IR evaluation and other related fields, providing new paths for semantic enrichment of the experimental data. Furthermore, we shall extend the DIRECT provenance event section by keeping track the role and the groups to which a user belonged where a specific action on a resource was taken.

DIRECT RDF model can also play a significant role in the call for better transparency and reproducibility in science (Baggerly, 2010). Indeed, it can be paired up with data citation methodologies (Buneman et al., 2016; Buneman and Silvello, 2010; Pröll and Rauber, 2015) in order to define a methodology to connect results in scientific papers with the actual data on which they are based as well as to sustain scientific claims as proposed in (Silvello, 2015).

Additionally, we plan to improve the automatically constructed taxonomy used in this work by making use of hierarchical relations provided in the DBpedia category structure and by using a disambiguation approach for grounding expertise topics.

Lastly, when it will arise, we plan to tackle the problem of name entity disambiguation as the dataset grows and the number of users (i.e., contribution authors) expands with the use of the dataset. Indeed, this issue does not impact the current dataset given the relatively small size of the IR community we consider here, but it has to be taken into account if we enlarge the boundaries of the system.

Ackwnoledgements

This work has been funded in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (INSIGHT). The authors would like to thank Barry Coughlan for his suggestions and collaboration.

References

Ackoff, R. L. (1989). From Data to Wisdom. *Journal of Applied Systems Analysis*, 16:3–9.

Agirre, E., Di Nunzio, G. M., Ferro, N., Mandl, T., and Peters, C. (2009). CLEF 2008: Ad Hoc Track Overview. In Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G. J. F., Kurimo, M., Mandl, T., and Peñas, A., editors, *Evaluating Systems for Multilingual and Multimodal Information Access: Ninth Workshop of the Cross–Language Evaluation Forum (CLEF 2008). Revised Selected Papers*, pages 15–37. Lecture Notes in Computer Science (LNCS) 5706, Springer, Heidelberg, Germany.

Agosti, M., Berendsen, R., Bogers, T., Braschler, M., Buitelaar, P., Choukri, K., Di Nunzio, G. M., Ferro, N., Forner, P., Hanbury, A., Friberg Heppin, K., Hansen, P., Järvelin, A., Larsen, B., Lupu, M., Masiero, I., Müller, H., Peruzzo, S., Petras, V., Piroi, F., de Rijke, M., Santucci, G., Silvello, G., and Toms, E. (2012a). PROMISE Retreat Report Prospects and Opportunities for Information Access Evaluation. *SIGIR Forum*, 46(2):60–84.

Agosti, M., Di Buccio, E., Ferro, N., Masiero, I., Peruzzo, S., and Silvello, G. (2012b). DIRECTions: Design and Specication of an IR Evaluation Infrastructure. In Catarci, T., Forner, P., Hiemstra, D.,

Peñas, A., and Santucci, G., editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics. Proceedings of the Third International Conference of the CLEF Initiative (CLEF 2012)*, pages 88–99. Lecture Notes in Computer Science (LNCS) 7488, Springer, Heidelberg, Germany.

Agosti, M., Di Nunzio, G. M., and Ferro, N. (2007a). 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., and Stempfhuber, M., editors, *Evaluation of Multilingual and Multi-modal Information Retrieval : Seventh Workshop of the Cross–Language Evaluation Forum (CLEF 2006). Revised Selected Papers*, pages 11–20. Lecture Notes in Computer Science (LNCS) 4730, Springer, Heidelberg, Germany.

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

Agosti, M. and Ferro, N. (2009). Towards an Evaluation Infrastructure for DL Performance Evaluation. In Tsakonas, G. and Papatheodorou, C., editors, *Evaluation of Digital Libraries: An insight into useful applications and methods*, pages 93–120. Chandos Publishing, Oxford, UK.

Agosti, M., Ferro, N., Fox, E. A., and Gonçalves, M. A. (2007c). Modelling DL Quality: a Comparison between Approaches: the DELOS Reference Model and the 5S Model. In Thanos, C., Borri, F., and Launaro, A., editors, *Second DELOS Conference - Working Notes*. ISTI-CNR, Gruppo ALI, Pisa, Italy.

Agosti, M., Ferro, N., and Silvello, G. (2016). Digital Library Interoperability at High Level of Abstraction. *Future Generation Computer Systems*, 55:129–146.

Agosti, M., Ferro, N., and 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, A. P., and Wenfei, F., editors, *Proc. 20th Int. Conference on Information and Knowledge Management (CIKM 2011)*, pages 2631–2632. ACM Press, New York, USA.

Allan, J., Aslam, J., Azzopardi, L., Belkin, N., Borlund, P., Bruza, P., Callan, J., Carman, M. Clarke, C., Craswell, N., Croft, W. B., Culpepper, J. S., 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, D. W., de Rijke, M., Sakai, T., Sanderson, M., Scholer, F., Si, L., Thom, J., Thomas, P., Trotman, A., Turpin, A., de Vries, A. P., Webber, W., Zhang, X., and 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.

Angelini, M., Ferro, N., Santucci, G., and Silvello, G. (2014). VIRTUE: A Visual Tool for Information Retrieval Performance Evaluation and Failure Analysis. *J. Vis. Lang. Comput.*, 25(4):394–413.

Angelini, M., Ferro, N., Santucci, G., and Silvello, G. (2016). A Visual Analytics Approach for What-If Analysis of Information Retrieval Systems. In Perego, R., Sebastiani, F., Aslam, J., Ruthven, I., and Zobel, J., editors, *Proc. 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016)*. ACM Press, New York, USA.

Arguello, J., Crane, M., Diaz, F., Lin, J., and Trotman, A. (2015). Report on the SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR). *SIGIR Forum*, 49(2).

Armstrong, T. G., Moffat, A., Webber, W., and Zobel, J. (2009). EvaluatIR: An Online Tool for Evaluating and Comparing IR Systems. In Allan, J., Aslam, J. A., Sanderson, M., Zhai, C., and Zobel, J., editors, *Proc. 32nd Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*, page 833. ACM Press, New York, USA.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley, Harlow, England.

Baggerly, K. (2010). Disclose all Data in Publications. *Nature*, (467):401.

Bailey, P., de Vries, A. P., Craswell, N., and Soboroff, I. (2007). Overview of the TREC 2007 Enterprise Track. In Voorhees, E. M. and Buckland, L. P., editors, *The Sixteenth Text REtrieval Conference Proc. (TREC 2007)*. National Institute of Standards and Technology (NIST), Special Pubblication 500-274, Washington, USA.

Balog, K., Azzopardi, L., and de Rijke, M. (2009). A language modeling framework for expert finding. *Information Processing & Management*, 45(1):1–19.

Balog, K., Bogers, T., Azzopardi, L., de Rijke, M., and van den Bosch, A. (2007). Broad expertise retrieval in sparse data environments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SI-

GIR '07, pages 551–558, New York, NY, USA. ACM.

Balog, K. and de Rijke, M. (2007). Determining Expert Profiles (With an Application to Expert Finding). In *Proc. of the International Joint Conferences on Artificial Intelligence (IJCAI 2007)*, pages 2657–2662, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Balog, K., de Rijke, M., and Azzopardi, L. (2006). Formal Models for Expert Finding in Enterprise Corpora. In *Proc. of the 29th annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '06*, pages 43–50, New York, New York, USA. ACM Press.

Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., and Si, L. (2012). Expertise Retrieval. *Foundations and Trends in Information Retrieval (FnTIR)*, 6(2-3):127–256.

Batini, C. and Scannapieco, M. (2006). *Data Quality. Concepts, Methodologies and Techniques*. Springer-Verlag, Heidelberg, Germany.

Berendsen, R., Balog, K., Bogers, T., van den Bosch, A., and de Rijke, M. (2013). On the Assessment of Expertise Profiles. *Journal of the American Society for Information Science and Technology (JASIST)*, 64(10):2024–2044.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22.

Bordea, G., Bogers, T., and Buitelaar, P. (2013a). Benchmarking Domain-Specific Expert Search Using Workshop Program Committees. In *Workshop on Computational Scientometrics: Theory and Applications, at CIKM*.

Bordea, G., Kirrane, S., Buitelaar, P., and Pereira, B. O. (2012). Expertise Mining for Enterprise Content Management. In Calzolari, N., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proc. of the Eighth Int. Conference on Language Resources and Evaluation (LREC-2012)*, pages 3495–3498. European Language Resources Association (ELRA).

Bordea, G., Polajnar, T., and Buitelaar, P. (2013b). Domain-Independent Term Extraction Through Domain Modelling. In *10th International Conference on Terminology and Artificial Intelligence*.

Borgman, C. L. (2012). The Conundrum of Sharing Research Data. *JASIST*, 63(6):1059–1078.

Borgman, C. L. (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.

Buneman, P. (2013). The providence of provenance. In Gottlob, G., Grasso, G., Olteanu, D., and Schallhart, C., editors, *Proc. of the 29th British National Conference on Databases, BNCOD 2013*, volume 7968 of *Lecture Notes in Computer Science*, pages 7–12. Springer.

Buneman, P., Davidson, S. B., and Frew, J. (2016). Why data citation is a computational problem. *Communications of the ACM (CACM)*, forthcoming.

Buneman, P. and Silvello, G. (2010). A Rule-Based Citation System for Structured and Evolving Datasets. *IEEE Data Eng. Bull.*, 33(3):33–41.

Burnett, S., Clarke, S., Davis, M., Edwards, R., and Kellett, A. (2006). *Enterprise Search and Retrieval. Unlocking the Organisation's Potential*. Butler Direct Limited.

Campbell, C. S., Maglio, P. P., Cozzi, A., and Dom, B. (2003). Expertise identification using email communications. In *CIKM '03: Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pages 528–531, New Orleans, LA.

Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobreva, M., Katifori, V., and Schuldt, H. (2007). *The DELOS Digital Library Reference Model. Foundations for Digital Libraries*. ISTI-CNR at Gruppo ALI, Pisa, Italy, `http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf`.

Candela, L., Castelli, D., Manghi, P., and Tani, A. (2015). Data Journals: A Survey. *Journal of the Association for Information Science and Technology*, page IN PRINT.

Chapelle, O., Ji, S., Liao, C., Velipasaoglu, E., Lai, L., and Wu, S.-L. (2011). Intent-Based Diversification of Web Search Results: Metrics and Algorithms. *Inf. Retr.*, 14(6):572–592.

Cheney, J., Chiticariu, L., and Tan, W. C. (2009). Provenance in Databases: Why, How, and Where. *Foundations and Trends in Databases*, 1(4):379–474.

Cleverdon, C. W. (1997). The Cranfield Tests on Index Languages Devices. In Spärck Jones, K. and Willett, P., editors, *Readings in Information Retrieval*, pages 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA.

Croft, W. B., Metzler, D., and Strohman, T. (2009). *Search Engines: Information Retrieval in Practice*. Addison-Wesley, Reading (MA), USA.

Demartini, G. (2007). Finding experts using wikipedia. In *Proceedings of the Workshop on Finding Experts on the Web with Semantics (FEWS2007) at ISWC/ASWC2007*, pages 33–41.

Di Buccio, E., Di Nunzio, G. M., and Silvello, G. (2013). A Curated and Evolving Linguistic Linked Dataset. *Semantic Web*, 4(3):265–270.

Draganidis, F. and Metzas, G. (2006). Competency based management: A review of systems and approaches. *Information Management and Computer Security*, 14(1):51–64.

Dussin, M. and Ferro, N. (2009). Managing the Knowledge Creation Process of Large-Scale Evaluation Campaigns. In Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., and Tsakonas, G., editors, *Proc. 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*, pages 63–74. Lecture Notes in Computer Science (LNCS) 5714, Springer, Heidelberg, Germany.

Ferro, N. (2014). CLEF 15th Birthday: Past, Present, and Future. *SIGIR Forum*, 48(2):31–55.

Ferro, N., Hanbury, A., Müller, H., and 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.

Ferro, N. and Silvello, G. (2014a). CLEF 15th Birthday: What Can We Learn From Ad Hoc Retrieval? In Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., and Toms, E., editors, *Information Access Evaluation – Multilinguality, Multimodality, and Interaction. Proc. of the Fifth Int. Conference of the CLEF Initiative (CLEF 2014)*, pages 31–43. Lecture Notes in Computer Science (LNCS) 8685, Springer, Heidelberg, Germany.

Ferro, N. and Silvello, G. (2014b). Making it Easier to Discover, Re-Use and Understand Search Engine Experimental Evaluation Data. *ERCIM News*, 96:26–27.

Ferro, N. and Silvello, G. (2015). Rank-Biased Precision Reloaded: Reproducibility and Generalization. In Fuhr, N., Rauber, A., Kazai, G., and Hanbury, A., editors, *Advances in Information Retrieval. Proc. 37th European Conference on IR Research (ECIR 2015)*, pages 768–780. Lecture Notes in Computer Science (LNCS) 9022, Springer, Heidelberg, Germany.

Ferro, N. and Silvello, G. (2016). A General Linear Mixed Models Approach to Study System Component Effects. In Perego, R., Sebastiani, F., Aslam, J., Ruthven, I., and Zobel, J., editors, *Proc. 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016)*. ACM Press, New York, USA.

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

Fricke, M. (2009). The Knowledge Pyramid: a Critique of the DIKW Hierarchy. *Journal of Information Science*, 35(2):131–142.

Gollub, T., Stein, B., Burrows, S., and Hoppe, D. (2012). TIRA: Configuring, Executing, and Disseminating Information Retrieval Experiments. In Hameurlain, A., Tjoa, A. M., and Wagner, R., editors, *23rd International Workshop on Database and Expert Systems Applications, DEXA 2012*, pages 151–155. IEEE Computer Society.

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

Harman, D. K. (2011). *Information Retrieval Evaluation*. Morgan & Claypool Publishers, USA.

Harman, D. K., Braschler, M., Hess, M., Kluck, M., Peters, C., Schaüble, P., and Sheridan, P. (2001). CLIR Evaluation at TREC. In Peters, C., editor, *Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum (CLEF 2000)*, pages 7–23. Lecture Notes in Computer Science (LNCS) 2069, Springer, Heidelberg, Germany.

Harman, D. K. and Voorhees, E. M., editors (2005). *TREC. Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge (MA), USA.

Heath, T. and 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.

Hennicke, S., Olensky, M., de Boer, V., Isaac, A., and Wielemaker, J. (2011). Conversion of EAD into EDM Linked Data. In Prediu, L., Hennicke, S., Nürnberger, A., Mitschick, A., and Ross, S., editors, *Proc. 1st International Workshop on Semantic Digital Archives (SDA 2011)* `http://ceur-ws.org/Vol-801/`, pages 82–88.

Hersey, A., Senger, S., and Overington, J. P. (2012). Open Data for Drug Discovery: Learning from the Biological Community. *Future Med. Chem.*, 4(15):1865–1867.

Hooper, C. J., Marie, N., and Kalampokis, E. (2012). Dissecting the butterfly: representation of disciplines publishing at the web science conference series. In Contractor, N. S., Uzzi, B., Macy, M. W., and Nejdl, W., editors, *WebSci*, pages 137–140. ACM.

Isaac, A. and Haslhofer, B. (2013). Europeana Linked Open Data - data.europeana.eu. *Semantic Web*,

4(3):291–297.

ISO 9000 (2005). Quality management systems – Fundamentals and vocabulary. Recommendation ISO 9000:2005.

Kharazmi, S., Scholer, F., Vallet, D., and Sanderson, M. (2016). Examining Additivity and Weak Baselines. *ACM Transactions on Information Systems (TOIS)*.

Lagoze, C., Van De Sompel, H., Johnston, P., Nelson, M., Sanderson, R., and Warner, S. (2008). ORE Specification – Abstract Data Model – Version 1.00. `http://www.openarchives.org/ore/1.0/datamodel`.

Leidig, J. P. (2012). *Epidemiology Experimentation and Simulation Management through Scientific Digital Libraries*. PhD thesis, Virginia Tech.

Lupu, M. and Hanbury, A. (2013). Patent Retrieval. *Foundations and Trends in Information Retrieval (FnTIR)*, 7(1):1–97.

Macdonald, C. and Ounis, I. (2006). Voting for candidates: adapting data fusion techniques for an expert search task. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 387–396, New York, NY, USA. ACM.

Maybury, M. (2006). Expert finding systems. Technical Report MTR 06B000040, MITRE Corporation.

Mimno, D. and McCallum, A. (2007). Expertise Modeling for Matching Papers with Reviewers. In *SIGKDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 500–509.

Müller, H. (2013). Medical (Visual) Information Retrieval. In Agosti, M., Ferro, N., Forner, P., Müller, H., and Santucci, G., editors, *Information Retrieval Meets Information Visualization – PROMISE Winter School 2012, Revised Tutorial Lectures*, pages 155–166. Lecture Notes in Computer Science (LNCS) 7757, Springer, Heidelberg, Germany.

Ngonga Ngomo, A.-C. (2012). On link discovery using a hybrid approach. *Journal on Data Semantics*, 1(4):203–217.

Petkova, D. and Croft, W. B. (2006). Hierarchical language models for expert finding in enterprise corpora. In *Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence*, ICTAI '06, pages 599–608, Washington, DC, USA. IEEE Computer Society.

Pröll, S. and Rauber, A. (2015). Asking the Right Questions - Query-Based Data Citation to Precisely Identify Subsets of Data. *ERCIM News*, (100).

Robertson, S. E. (2008). On the history of evaluation in IR. *Journal of Information Science*, 34(4):439–456.

Rodriguez, M. A. and Bollen, J. (2008). An Algorithm to Determine Peer-Reviewers. In *'08: Proceedings of the Seventeenth International Conference on Information and Knowledge Management*, pages 319–328. ACM.

Rowe, B. R., Wood, D. W., Link, A. L., and Simoni, D. A. (2010). *Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program*. RTI Project Number 0211875, RTI International, USA. `http://trec.nist.gov/pubs/2010.economic.impact.pdf`.

Rowley, J. (2007). The Wisdom Hierarchy: Representations of the DIKW Hierarchy. *Journal of Information Science*, 33(2):163–180.

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, USA.

Serdyukov, P., Taylor, M., Vinay, V., Richardson, M., and White, R. (2011). Automatic people tagging for expertise profiling in the enterprise. In Clough, P., Foley, C., Gurrin, C., Jones, G., Kraaij, W., Lee, H., and Mudoch, V., editors, *Advances in Information Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pages 399–410. Springer Berlin Heidelberg.

Silvello, G. (2015). A Methodology for Citing Linked Open Data Subsets. *D-Lib Magazine*, 21(1/2).

Soboroff, I., de Vries, A. P., and Craswell, N. (2007). Overview of the trec 2006 enterprise track. In *The fifteenth Text REtrieval Conference Proceedings (TREC 2006)*.

Stasinopoulou, T., Bountouri, L., Kakali, C., Lourdi, I., Papatheodorou, C., Doerr, M., and Gergatsoulis, M. (2007). Ontology-Based Metadata Integration in the Cultural Heritage Domain. In Goh, D., Cao, T., Sølvberg, I. T., and Rasmussen, E., editors, *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, volume 4822 of *Lecture Notes in Computer Science*, pages 165–175. Springer Berlin Heidelberg.

Tanaka, J. W. and Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23(3):457–482.

Thiagarajan, R., Manjunath, G., and Stumptner, M. (2008). *Finding experts by semantic matching of user profiles*. PhD thesis, CEUR-WS.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth.

W3C (2004a). RDF Primer – W3C Recommendation 10 February 2004.

W3C (2004b). Resource Description Framework (RDF): Concepts and Abstract Syntax – W3C Recommendation 10 February 2004.

W3C (2009a). SKOS Simple Knowledge Organization System Primer – W3C Working Group Note 18 August 2009.

W3C (2009b). SKOS Simple Knowledge Organization System Reference – W3C Recommendation 18 August 2009.

Zapilko, B., Schaible, J., Mayr, P., and Mathiak, B. (2013). TheSoz: A SKOS representation of the thesaurus for the social sciences. *Semantic Web*, 4(3):257–263.

Zeleny, M. (1987). Management Support Systems: Towards Integrated Knowledge Management. *Human Systems Management*, 7(1):59–70.

Zobel, J., Webber, W., Sanderson, M., and Moffat, A. (2011). Principles for Robust Evaluation Infrastructure. In Agosti, M., Ferro, N., and Thanos, C., editors, *Proc. Workshop on Data infrastructurEs for Supporting Information Retrieval Evaluation (DESIRE 2011)*, pages 3–6. ACM Press, New York, USA.