IR Scientific Data: How to Semantically Represent and Enrich Them

Extended abstract of (Silvello et al., 2016)

Toine Bogers
Aalborg University Copenhagen, Denmark
toine@hum.aau.dk

Georgeta Bordea **Paul Buitelaar**
Insight Centre, National University of Ireland, Galway, Ireland
{georgeta.bordea, paul.buitelaar}@insight-centre.org

Nicola Ferro **Gianmaria Silvello**
University of Padua, Padua, Italy
{ferro, silvello}@dei.unipd.it

Abstract

English. 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. We discuss how to annotate and interlink this data, by proposing a method 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.

Italiano. *La valutazione sperimentale condotta mediante campagne internazionali su larga scala, è un pilastro fondante dello sviluppo scientifico e dell'avanzamento tecnologico dei sistemi di reperimento dell'informazione. Queste attività di valutazione producono una grande quantità di dati sperimentali che costituiscono la base per la conseguente produzione scientifica e lo sviluppo di nuovi sistemi. In questo lavoro, si discute come annotare e collegare questi dati, proponendo un metodo per esporre i dati sperimentali come LOD nel Web e per usare tali dati come base per arricchirli. In questo contesto, viene proposto un approccio centrato sui topic per la ricerca di esperti, che affronta il problema dell'estrazione dei topic e il collegamento di questi con la "LOD cloud" e con i dati sperimentali.*

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 (Ferro et al., 2016; Silvello and Ferro, 2016).

Nevertheless, 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 (Harman, 2011), the Linked Open Data (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.

Therefore, the main contributions of this paper are to:

- define an Resource Description Framework (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.

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

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 1 shows an RDF graph, which provides a visual representation 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 supported by the Distributed Information Retrieval Evaluation Campaign Tool (DIRECT) system which provides the conceptual model for representing and enriching the data (Agosti and Ferro, 2009; Agosti et al., 2012).

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. 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 1, 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 1, 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[1] 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[2] linked open dataset.

In Figure 1 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

[1] `http://www.dbpedia.org/`
[2] `http://dblp.l3s.de/`

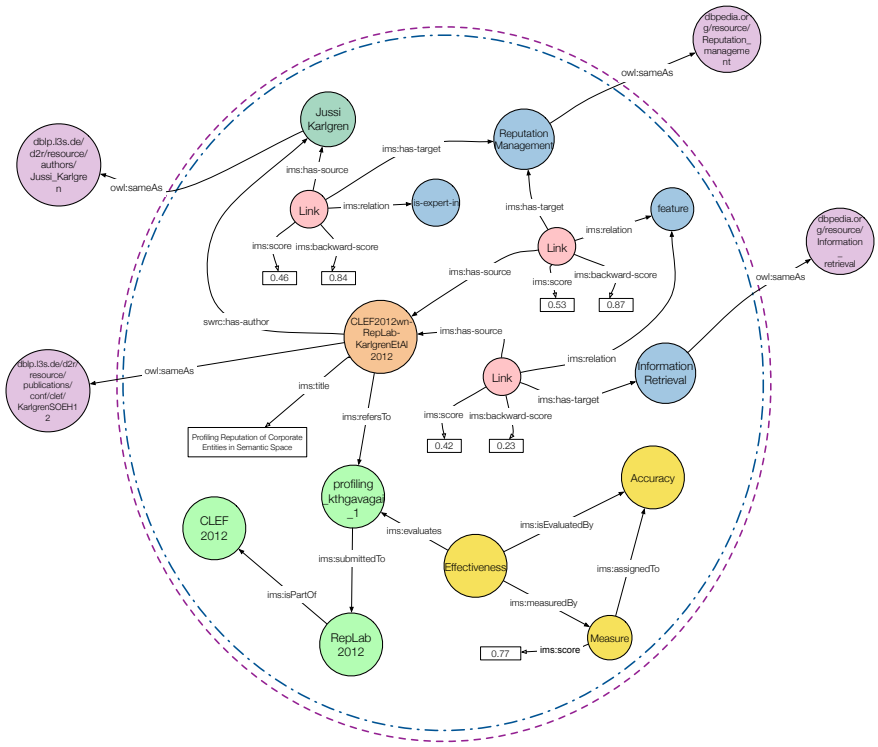

Figure 1: An example of RDF graph showing how expertise topics and expert profiles are used for enriching IR experimental data.

results of the evaluation activities to the claims reported by the contributions.

The details of the full RDF model are reported in (Silvello et al., 2016).

2.1 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 Conference and Labs of the Evaluation Forum (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).

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 1 with identifier `CLEF2012wn-RepLab-KarlgrenEt2012b` is:

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

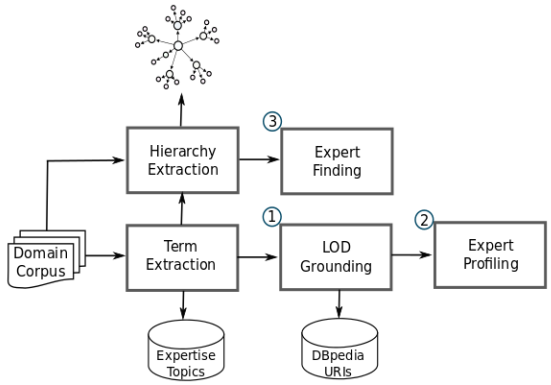

Figure 2: Data flow of the semantic enrichment approach

3 Semantic Enrichment

In this section we describe SOME methods for semantically enriching experimental IR data modelled as described above, by analysing unstructured data available in scientific publications. Figure 2 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. Then, we present a preliminary approach for enriching expertise topics by grounding them in the LOD cloud.

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). 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., 2013) and then noun phrases that include words from the domain model or that appear in their immediate context are selected as candidates.

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 overall 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 1: 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

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). Additional background knowledge, as found on the LOD cloud, can inform expert search at different stages.

A first step in the direction of exploiting this potential is to provide an entry point in the LOD cloud through DBpedia[3]. 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 above. Where available, concept descriptions are collected as well and used in our system.

Two approaches for grounding expertise topics on DBpedia have been evaluated. 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. 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.

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 1. Surprisingly, using lemmatization achieves a lower recall but higher precision but this might be due to the small size of the dataset.

Expert finding is the task of identifying the most

[3]DBpedia: `http://dbpedia.org/`

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. 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 2. The expert finding methods evaluated in this section include Experience (E), Relevance and Experience (RE) and Relevance, Experience and Area Coverage (REC).

Experience (E) is based on the idea that documents written by a person can be used as an indirect evidence of expertise, assuming that an expert often mentions his areas of interest. Relevance and Experience (RE) exploits the idea that 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. Relevance and expertise measure different aspects of expertise and can be combined to take advantage of both features. 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. 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* (REC) that measures whether an expert has in depth knowledge of an expertise topic, using an automatically constructed topical hierarchy.

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). Figure 3 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".

The details on the algorithms and weighting schemes for topic extraction, expert profiling, and expert finding are reported in (Silvello et al., 2016).

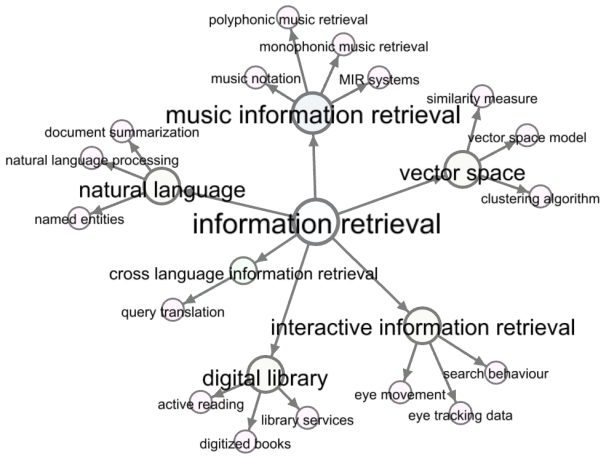

Figure 3: Sample hierarchical relations for the IR domain

4 Conclusion

In this paper we discussed the data modelling and the semantic enrichment of IR experimental data, as produced by large-scale evaluation campaigns.

In particular, the main results 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;
- 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 Digital Library (DL) methodologies for data access and management which in-

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

creasingly exploit the LOD paradigm (Hennicke et al., 2011; Di Buccio et al., 2013). 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.

References

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

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

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

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

C. L. Borgman. 2015. *Big Data, Little Data, No Data*. MIT Press.

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

N. Ferro, N. Fuhr, K. Järvelin, N. Kando, M. Lippold, and J. Zobel. 2016. Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science". *SIGIR Forum*, 50(1):68–82, June.

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

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

Clare J. Hooper, Nicolas Marie, and Evangelos Kalampokis. 2012. Dissecting the butterfly: representation of disciplines publishing at the web science conference series. In Noshir S. Contractor, Brian Uzzi, Michael W. Macy, and Wolfgang Nejdl, editors, *WebSci*, pages 137–140. ACM.

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

G. Silvello and N. Ferro. 2016. "Data Citation is Coming". Introduction to the Special Issue on Data Citation. *Bulletin of IEEE Technical Committee on Digital Libraries (IEEE-TCDL)*, 12(1):1–5, May.

G. Silvello, G. Bordea, N. Ferro, P. Buitelaar, and T. Bogers. 2016. Semantic Representation and Enrichment of Information Retrieval Experimental Data. *International Journal on Digital Libraries (IJDL)*.

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