Published in Proceedings of the 7th Metadata and Semantics Research Conference, Thessaloniki, Greece, 19th-22nd November 2013. Springer Communications in Computer and Information Science. P. 60-66, which should be cited to refer to this work.

Linking Search Results, Bibliographical Ontologies and Linked Open Data Resources

Fabio Ricci, Javier Belmonte, Eliane Blumer, René Schneider

Haute Ecole de Gestion de Genève, 7 route de Drize, CH-1227 Carouge
{fabio.fr.ricci, javier.belmonte, eliane.blumer, rene.schneider}@hesge.ch

Abstract. This paper describes a lightweight approach to build an environment for scientific research that connects user-selected information resources with domain specific ontologies and the linked open data cloud. Search results are converted into RDF triples to match with ontology subjects in order to derive relevant subjects and to find related documents in external repositories data that are stored in the Linked Open Data Cloud. With the help of this deterministic algorithm for analyzing and ranking search subjects, the explicit searching process, as effectuated by the user, is implicitly supported by the LOD-technology.

Keywords: Innovative Scientific Search, Metadata Reusability, Linked Open Data Technologies

1 Introduction

Libraries have always been interested in developing meta data descriptions for the documents they take care of. In recent years, more and more of these taxonomies and the thesauri developed for this purpose are converted into ontologies or ontology-like repositories (i.e. the data is expressed as ontologies are) that can be used to support scientific search in user created search environments [1]. Users can expand or narrow their search results with the help of the ontology terms that are presented in the faceted browsing menu and improve their search. Yet, this search process has to be triggered by the user who generally seems to prefer to use simple and fast search environments that are easy to understand and do not need prior explanations. Alternatively, the search topic, search results and the ontological terms can be combined and connected in a kind of black box. In this context we follow the berrypicking metaphor described by Bates in [2] where searching is not seen as a linear process, but a meandering way finding process. In our system RODIN (=ROue D'INformation, i.e. information wheel), we developed an interface

that enables the user to explicitly perform scientific search by picking search terms from ontologies and search results. Due to the complexity of the system, we tried to find a solution that makes parts of the berrypicking process implicit by the help of Linked Open Data Technology, as described in the following paper.

2 Context

2.1 Prior Work

RODIN is a personalizable information portal that relies on the Posh Portal (http://sourceforge.net/projects/posh Portaneo) for widget administration. Widgets operations are carried on by our ad-hoc developed object oriented framework, which easily allows the integration of new information sources into widgets. The user selects data sources and runs a distributed search with the results being displayed in each widget and stored in a database in a homogenized format for fast reuse. This framework has been extended by adding the following components: a) ontological facets, i.e. RDF thesauri based on the SKOS (=Simple Knowledge Organization System) [3] model, b) a SOLR (=Searching On Lucene with Index Replication) index machine for fast information processing concerning widget results using the vector space model document metrics, result similarity functions, term distance measures representing the vector space distance between all documents, term matching inside RDF thesauri and ranking methods, c) an interactive graphical visualization of the SKOS part of DBPedia [4] and STW (=Standard Thesaurus Wirtschaftswissenschaften, i.e. the standard thesaurus for economic sciences) graph, d) thesauri based on SKOS enabling navigable auto complete suggestions.

Afterwards and as described in this paper, this architecture was enhanced by an RDF engine that enriches RODIN search results with external LOD documents relying on shared subjects as described in detail in the following part. One reason for this extension lied in a shift of the data layer from a relational database to a triple store, enhancing the compatibility of the results with current further LOD sources. Search results as well as subjects and result-related information are stored as RDF generic resources and made available (querieable) through a further LOD interface called dbRODIN (with db for database). We called this operation “RDFization”. The other purpose is

in enhancing the number of the attached thesauri and enhancing the power of the filtering functionality while augmenting the usability of the same system.

2.2 Related work

Compared to other work done in this domain, RODIN tries to find new ways to build bibliographical search engines by subscribing to the view of the web of data, similar to the Europeana approach [5] without necessarily relying on the FRBR (= Functional Requirements for Bibliographical records) concept [6]. RODIN tries to balance the information seeking and management process between the user and the machine as well as between internal and external information resources. This means that only part of the information seeking process is in the hand of the user and some parts are taken over by the retrieval engine. Only a very specific part of the information processed is kept in internal representations or repositories. The balance has to be found between keeping large parts of the search space external and without prior processing such as indexing or harvesting [7] and by adequately building own information repositories that are linkable to other open data available online.

3 Linking Documents, Ontologies and Linked Open Data

3.1 Rodin Architecture

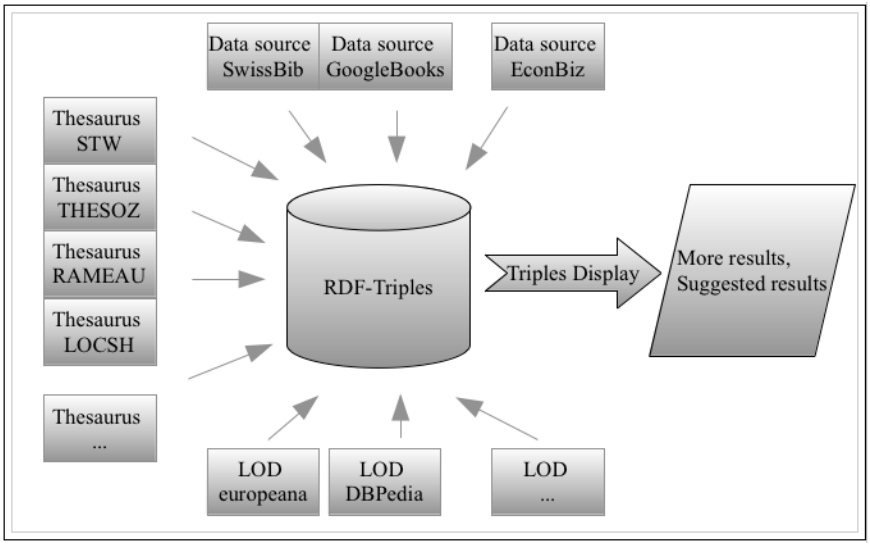

Fig. 1. RODIN general architecture based on widget data, thesauri and LOD sources

As stated before, RODIN (see Fig. 1.) integrates already a number of information sources as widgets (e.g. Google Books, SwissBib, and EconBiz) and several RDF thesauri (e.g. DBPedia, STW, THESOZ (= Thesaurus for the Social Sciences), RAMEAU (=Répertoire d'autorité-matière encyclopédique et alphabétique unifié, i.e. the French standard for subject headings), and LCSH (= Library of Congress Subject Headings)) for the semantic expansion of search related terms.

A significant enhancement of RODIN's result delivery coverage comes from an added RDF processing component, which accesses further related documents from LOD data sources based on shared subjects extracted from widget result documents. This added RDF module transforms widget result information and subjects into an own homogenous RDF store. All further operations, semantic expansion and ranking of subjects, and imported LOD documents are performed on the basis of the information in the RDF store. During result RDFization, widget result documents are processed as follows:

1. Subject expansion: Every subject provided from a specific widget document is expanded with respect to configured RODIN ontological sources (thesauri) as activated by the user, which produces SKOS related subjects in the same language as the search term to be added to the subjects of the widget documents. Every related subject is stored in RODIN's RDF store.
2. Document expansion: Based on every original and related subject, documents in the same language as detected on the search term are retrieved from the LOD data sources and homogenized inside RODIN's RDF store.
3. RDF mirror: Besides the presented RDF process we created in RODIN the RDF mirror of its search data as well as the search results and the expanded data - called dbRODIN – offering a DBPedia like RDF graph navigator and an LOD SPARQL (=Simple Protocol and RDF Query Language) endpoint for public search results access and RDF download [8][9]. In this way RODIN gets its own LOD cloud which can be made accessible for shared use in the linked open data space.

3.2 RDFization and Linking of Documents

Figure 2 illustrates the steps in RODIN's RDFization process. In Step 1 a reference corpus for later subject ranking is built out of the search term and

matched against the connected RDF thesauri (Example a). The same action is performed on the subjects of each widget result document: delivered subjects are matched against the RDF thesauri to gather further related subjects (steps 3 and 4, Example b). All subjects in 3 and 4 are ranked in step 5 with respect to their vector space distance to the reference corpus (Example c). Using the ranked subjects of step 5, we search for documents inside the LOD sources (steps 6a, 6b), afterwards we homogenize the triples delivered by the LOD source (since we do not yet have multigraph processing RDF engines) on the given subjects and add them to RODIN's RDF store (Example d). Note that fetching LOD records delivers not only documents but also further related subjects. The latter will be added to RODIN's store as well. The quality or added value of suggested LOD documents relies strongly on how precisely subjects can be extracted from widget results. Missing subjects in widget result meta data force RODIN to "guess" subjects using the result title – this being a source of "noise" inside the illustrated RDFization process.

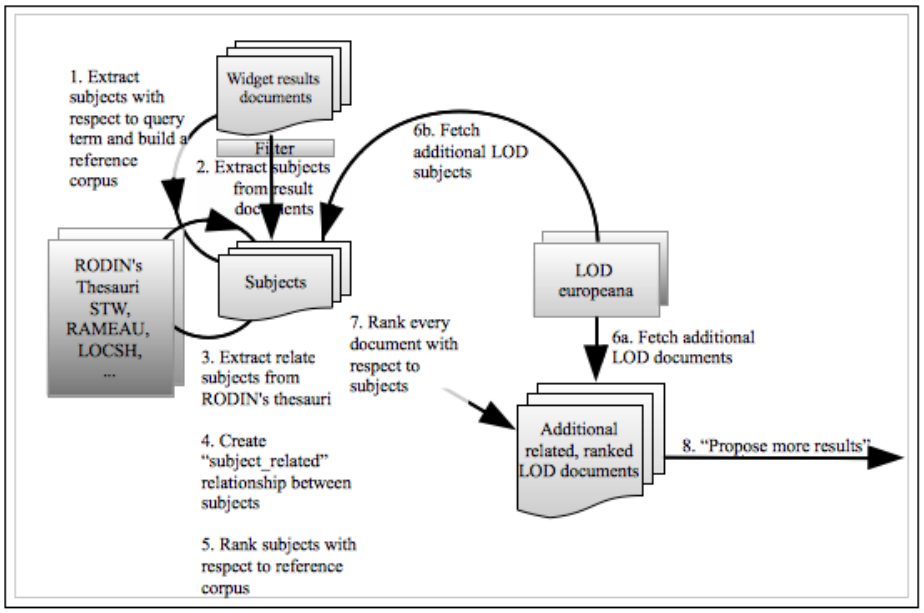

Fig. 2. RDFization Process in Rodin

The final step consists in ranking all the subjects again. All documents (widget documents and LOD documents) are re-ranked with respect to their (ranked) subjects (step 7). The order obtained (top ranked first) is used to display all found documents in RODIN's unified view from the most ranked

to the least ranked document. The following series of examples relies on the search term “digital economy” and shall be taken to follow the RDFization process in all his steps.

- Example a: Title is “digital economy and e-commerce technology“, derived subjects are: “digital, economy, e-commerce”.
- Example b: The reference corpus for “digital economy” built in RODIN is “digital economy, e-health, business intelligence system”.
- Example c: Ranked subject with respect to reference corpus are: “digital, economy, e-commerce, editorial, data protection, open source”.
- Example d: Homogenized LOD triplets on subject “e-commerce” are shown in table 1 – (shortened example).

epp:OID	dce:title	"Buying Online: Sequential Decision Making by Shopbot Visitors"
epp:OID	dce:description	"Forschungsbericht"
epp:OID:	dce:description	OID "Abstract: In this article we propose a two stage procedure to model demand decisions by customers who are balancing …"
epp:OID	dce:creator	"Winter-Ebmer, Rudolf"
epp:OID	dce:subject	"Ökonomie"
epp:OID	dce:subject	"Decision theory"
epp:OID	dce:subject	"**E-commerce**"
epp:OID	dce:subject	“Price comparison"
epp:OID	dce:subject	"Heuristics"
epp:OID	dce:type	"Text"
epp:OID	dce:date	"2008"
epp:OID	dce:language	"englisch"
epp:OID	dce:publisher	"Wien"
epp:OID	dce:publisher	OID "Kunst, Robert M. (Ed.) ; Fisher, Walter (Assoc. Ed.) ; Ritzberger, Klaus (Assoc. Ed.)"
epp:OID	dct:tableOfContents	"from the Table of Contents: Introduction; A Decision Procedure; Data and Estimation Strategy; Empirical Results; Robustness"
epp:OID	dct:extent	"24 pp."
epp:OID	dce:identifier	"oai:at.europana-local: SHI/000000471088"
epp:OID	dct:isPartOf	"Economics Series"
epp:OID	dct:isPartOf	"Kunst, Robert M. (Ed.); Fisher, Walter (Assoc. Ed.) ; Ritzberger, Klaus (Assoc. Ed.)"
epp:OID	dct:isPartOf	OID "Institut für Höhere Studien; Reihe Ökonomie; 225"
epp:OID	dct:issued	"2008, September"

Table 1. Homogenized external LOD triples on subject “e-commerce” (excerpt)

3.3 Creating the LOD store in dbRODIN

Since triples are used to connote resources, it is important to assign unique identifiers for each created RDF resource generated from search results and their expansions. In dbRODIN resources we find objects concerning works, articles, publishers; it is therefore mandatory to generate for each of these

resources unique identifiers as they are imported from an external LOD source as well as from a widget. In dbRODIN, unique identifiers are generated by compressing the resource description (eliminating punctuation) and by limiting the resulting id length. The corresponding dbRODIN RDF store contains only RODIN's graphs. The combination of the "rodin" namespace and the unique id guarantees uniqueness inside dbRODIN graph. Finally, in order to assure compatibility for any further processing of dbRODIN's triples outside the own data store, triples components have to be shaped using adequate common vocabularies. In dbRODIN, we use – besides a few own "rodin" terms – standard vocabularies like DublinCore (http://purl.org/dc/elements/1.1/), dcterms (http://purl.org/dc/terms/), bibo (http://bibliontology.com/bibo/bibo.php), bio (http://vocab.org/bio/0.1/), foaf (http://xmlns.com/foaf/0.1/), rdf (http://www.w3.org/1999/02/22-rdf-syntax-ns#), and rdfs (http://www.w3.org/2000/01/rdf-schema).

3.4 Personal result filters

RODIN users can benefit from a final result biasing on their specific scientific interests (e.g. economical experts or medical doctors) they have. Using a simple vector space distance algorithm and two freely definable sets of words – a positive and a negative "resonance" set – defining which terms should have a higher resonance and which ones should be less important, the user gets finally results with a higher resonance first, while less important ones are ranked lower in the result list, according to the preferences and rejections defined before.

4 Conclusions

In this paper we described an approach to extend and re-rank search results by connecting RDFized search results with subjects derived from bibliographical ontologies as well as external documents from the Linked Open Data cloud. This approach integrates smoothly three layers of information: web documents from priory selected information resources, semantically rich information from thesauri that were converted into a semantic web compatible format as well as external data that found no prior consideration in the search process but semantically match to the original search structure. The newly added documents built the basis for opening the scientific perspective and

may also be of value for suggesting cooperation in scientifically based social networks.

The system generates its own LOD space for public access and offers the benefit of sharing enriched search results from information specialists in an "LOD-way". Personal search filters re-rank relevant results e.g. with respect to the professional group the user is belonging to. Through its LOD interface, RODIN opens up to the linked open data community by sharing searches and results done by information professionals and scientists.

5 References

1. Belmonte, J., Blumer, E., Ricci, F., Schneider, R.: RODIN – An E-Science Tool for Managing Information in the Web of Documents and the Web of Knowledge. E-Science and Information Management. pp. 4-12, Springer, Berlin (2012)

2. Bates, M.J.: The design of browsing and berrypicking techniques for the online search interface. Online Review, 13, 407-424 (1989).

3. Miles, A., Brickley, D., Matthews, B., Wilson, M.: SKOS Core Vocabulary Specification. International Conference on Dublin Core and Metadata Applications, pp. 3-10 (2005)

4. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus for a Web of Open Data. Springer Lecture Notes in Computer Science. Springer, Berlin, pp. 722-735 (2007)

5. Aloia, N., Concordia, C., Meghini, C.: „The Europeana Linked Open Data Pilot Server". *Digital Libraries and Archives*. Springer, Berlin, pp. 241-248 (2013)

6. Howarth, L. C.: FRBR and Linked Data: Connecting FRBR and Linked Data. Cataloging & Classification Quarterly, 50 (5-7), 763-776 (2012)

7. Introna, L, Nissenbaum, H.: Defining the Web: the politics of search engines. Computer 33 (1), pp. 54-62 (2000)

8. Sheth, A., Krishnaprasad, T.: "Semantics-empowered Web 3.0, Managing Enterprise, Social, Sensor, and Cloud based Data and Services for Advanced Applications". Synthesis Lectures on Data Management, 4 (6), pp. 1-175 (2012)

9. Fensel, D., Facca, F.M., Simperl, E., Toma, I.: Semantic Web Services. Springer, Berlin (2011)