The Semantic Lancet Project: A Linked Open Dataset for Scholarly Publishing

Andrea Bagnacani[1], Paolo Ciancarini[1,2], Angelo Di Iorio[1], Andrea Giovanni Nuzzolese[2], Silvio Peroni[1,2(✉)], and Fabio Vitali[1]

[1] Department of Computer Science and Engineering, University of Bologna, Bologna, Italy
andrea.bagnacani@studio.unibo.it,
{paolo.ciancarini,angelo.diiorio,silvio.peroni,fabio.vitali}@unibo.it
[2] Semantic Technology Laboratory, ISTC-CNR, Rome, Italy
andrea.nuzzolese@istc.cnr.it

Abstract. In this poster we introduce the *Semantic Lancet Project*, whose goal is to make available rich data about scholarly publications and to provide users with sophisticated services on top of those data.

Keywords: Data reengineering and enhancement · Linked open data · Scholarly data · Semantic lancet project · Semantic publishing · SPAR ontologies

1 Introduction

The availability of rich open (linked) data about scholarly data opens the way to novel applications for a large spectrum of users. The knowledge management of scholarly products is an emerging research area, and involves different users such as authors (for gathering personal repositories of papers), publishers (for constructing repositories of assets from venues), institutions and funding agencies (for ranking research assets). Even if there is interest in publishing such data as Linked Open Data (LOD), the current landscape is fragmented: some projects focus on bibliographic data (e.g., Nature Publishing Group LOD Platform[1]), others on authorship data (e.g., DBLP++[2]), others on citations (e.g., OpenCitation corpus [2,4]).

In this poster we introduce the *Semantic Lancet Project*, whose goal is to make available rich scholarly data and to provide users with sophisticated services on top of those data. The structure of the paper is as follows: Section 2 introduces our project, while Section 3 sketches out some future works.

[1] http://www.nature.com/developers/documentation/linked-data-platform/
[2] http://dblp.l3s.de/dblp++.php

P. Lambrix et al. (Eds.): EKAW 2014 Satellite Events, LNAI 8982, pp. 101–105, 2015.
DOI: 10.1007/978-3-319-17966-7_10

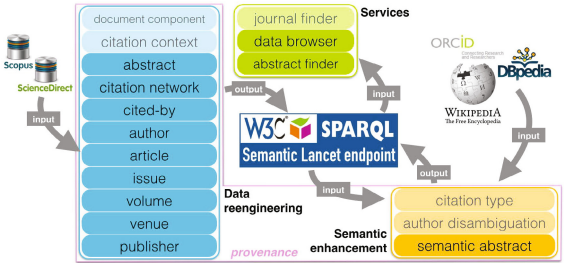

Fig. 1. The overall structure of the Semantic Lancet Project. The blurry blocks (e.g., *document component*, *citation type*, *journal finder*) are currently under development.

2 Semantic Lancet Project

The *Semantic Lancet Project* (http://www.semanticlancet.eu) is focused on building a Linked Open Dataset of scholarly publications. The aim of the project is twofold. On the one hand, we want to develop a series of scripts that allow us to produce proper RDF data compliant with the *Semantic Publishing and Referencing* (*SPAR*) *Ontologies* (http://www.sparontologies.net). On the other hand, we want to make publicly-available a rich RDF triplestore[3] (accompanied by a SPARQL endpoint) and a series of services built upon it starting from the data made available from Elsevier – while preparing the whole infrastructure in order to facilitate the future managing of data coming from other publishers. We have already converted and published on the Semantic Lancet triplestore all the data concerning the Journal of Web Semantics[4] (367 articles, 80920 RDF statements). The framework of the Semantic Lancet Project, summarised in Fig. 1, is composed basically by three macro sections, that we briefly introduce as follows[5].

Data reengineering. The data reengineering section is the one responsible of the translation of the raw data coming from the Science Direct and Scopus[6] repositories into RDF. Even if this could seem a duplication of the same data, gathering data from both repositories is a required action since some of the data in Scopus may be missing (e.g., there is no DOI specified for some articles) or wrong (e.g., the issue number of a certain article is zero) while they may be complete and correct in Science Direct, and vice versa. Such a 'twofold' approach improves the quality of the imported data[7].

[3] The triplestore we are currently using is Fuseki.
[4] http://www.journals.elsevier.com/journal-of-web-semantics
[5] The *provenance* module included in Fig. 1 is still under development.
[6] Science Direct (http://sciencedirect.com); Scopus (http://www.scopus.com).
[7] From a preliminary analysis performed considering 10 different journals coming from different academic disciplines, we decided to use Science Direct as base repository and Scopus for addressing incorrectness/missing in data.

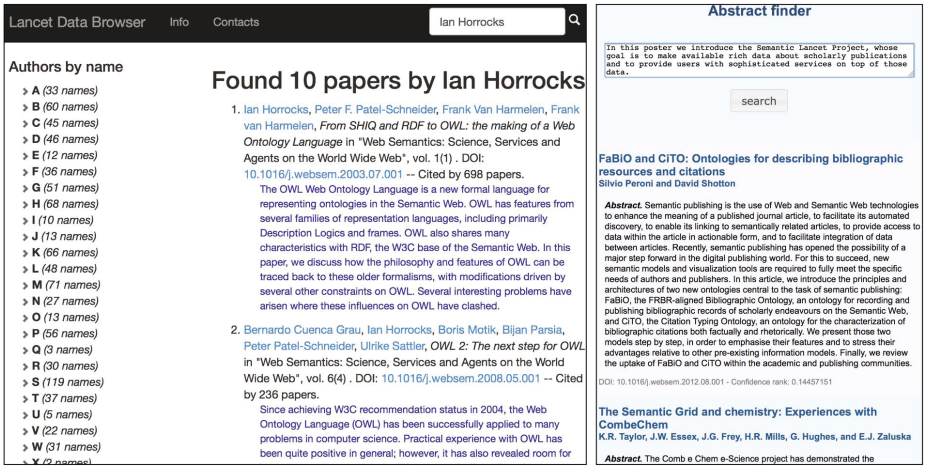

Fig. 2. The Semantic Lancet data browser (left) and abstract finder (right)

Basically, two kinds of data are requested by using the API made available by Elsevier[8]: those referring to metadata of articles and those concerning the full text of articles. The collected data are processed by a chain of scripts, one for each block of the data reengineering section shown in the left side of Fig. 1; each script retrieves all the data of interest and convert them into proper SPAR-based RDF statements.

Semantic enhancement. The following step of our workflow consists of enriching the dataset with more semantic data, that can be exploited for sophisticated end-user applications. On the one hand, some of these data are derived from the content of the papers by extracting semantic features from unstructured or semi-structured text and representing them as RDF [1]. On the other hand, the other data are the result of further refinement of the existing dataset.

Currently we have implemented a module for generating *semantic abstracts*. The generation of semantic abstracts relies mainly on FRED[9][3], which is a tool that implements deep machine reading methods based on Discourse Representation Theory, Linguistic Frames, and Ontology Design Patterns. FRED allows us to derive an OWL representation of the natural language sentences contained in the abstract, as well as to retrieve entities (e.g., DBpedia resources) that are cited in the abstracts.

Services. The framework is completed by a set of services for accessing, making sense and exploiting the (semantic) information available in the dataset. This part is shown in the top of Fig. 1 and consists of an extensible set of modules. Each module is independent from the others, it is built on top of the underlying

[8] http://www.developers.elsevier.com/devcms/content-apis
[9] http://wit.istc.cnr.it/stlab-tools/fred

semantically-enriched data, and it provides particular functionalities to the user. We have currently implemented two experimental modules, both available in the project website: the *data browser* and the *abstract finder*.

The *data browser* (cf. Fig. 2, on the left) is an interactive and user-friendly interface, with autocompletion and incremental loading of content, that allows users to easily access authors and their papers. The solution we propose is to hide the intrinsic complexities of the data and of the underlying technologies, giving users an higher-level view over the dataset content. The tool, in fact, does not show directly the entities stored in the dataset but groups those entities in more abstract "objects" that are finally shown to the users. A paper, for instance, is internally modelled according to the SPAR model, thus according to FRBR [5], and is defined in terms of Work, Expression(s), Manifestation(s) and Item(s). The dataset items are transparent to the user (though are available to software agents) who only deals with the concept of "paper". The same happens for properties: there is no distinction between object properties and data properties visible to the user. That distinction is in the dataset, and can be browsed on demand, but is fully hidden by default. Users, in fact, are not expected to master directly the Semantic Web technologies.

The *abstract finder* allows us to retrieve relevant papers according to their textual abstract as well as to the related *semantic abstract* – i.e., by exploiting the semantic information about concepts, events, roles and named entities produced during the *semantic enhancement* step. This tool works in two phases. First, it creates a *semiotic index* of the semantic abstracts with respect to the related taxonomy of types defined within them – that are aligned to WordNet synsets and DBpedia resources. In this way we can index the papers according to the textual content of their abstracts as well as the concepts represented in that content. Finally, a simple interface (cf. Fig. 2, on the right) allows users to query for papers having similar abstracts to the text specified as input.

3 Conclusions

In this poster we presented the *Semantic Lancet Project*, which aims at making available rich scholarly data and at providing users with sophisticated services on top of those data. There are other services we can build on our dataset in the future, once extended with more journals and kinds of data – e.g., citation contexts, citation functions, authors' affiliations and documents' internal components. In addition, some refinements are still under development, such as the disambiguation of authors' names and the inclusion of provenance information.

Acknowledgments. We would like to thank Elsevier for granting access to Scopus and ScienceDirect APIs.

References

1. Gangemi, A.: A comparison of knowledge extraction tools for the semantic web. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 351–366. Springer, Heidelberg (2013). doi:10.1007/978-3-642-38288-824
2. Peroni, S., Gray, T., Dutton, A., Shotton, D.: Setting our bibliographic references free: towards open citation data. Journal of Documentation **71**(2), (2015). doi:10.1108/JD-12-2013-0166
3. Presutti, V., Draicchio, F., Gangemi, A.: Knowledge extraction based on discourse representation theory and linguistic frames. In: ten Teije, A., Völker, J., Handschuh, S., Stuckenschmidt, H., d'Acquin, M., Nikolov, A., Aussenac-Gilles, N., Hernandez, N. (eds.) EKAW 2012. LNCS, vol. 7603, pp. 114–129. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33876-212
4. Shotton, D.: Open citations. Nature **502**(7471) (2013). doi:10.1038/502295a
5. IFLA Study Group on the FRBR. Functional Requirements for Bibliographic Records (2009). http://www.ifla.org/publications/functional-requirements-for-bibliographic-records