Towards a Linked Open Dataset for Scholarly Publishing: Semantic Lancet Project

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

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

Abstract. There is an ever increasing interest in publishing Linked Open Datasets about scientific papers. The current landscape is very fragmented: some projects focus on bibliographic data, others on authorship data, others on citations, and so on. The quality is also heterogeneous and the production and maintenance of such datasets is difficult and time-consuming.
In this paper we introduce the Semantic Lancet Project, whose goal is to make available rich semantic data about scholarly publications and to provide users with sophisticated services on top of those data.
We developed a chain of tools that produce high-quality data from multiple sources. It has been successfully used to produce a rich and freely available LOD, described here as well.

Keywords: Data Reengineering, Linked Open Dataset, Scholarly publications, Semantic Enhancement, Semantic Publishing

1 Introduction

The authors of a scientific paper can cite another paper for different reasons. A very common reason is that the cited paper has been useful to those authors, for instance because it has proposed a problem to investigate or a solution to analyse. Another common reason is that the cited paper contains information that is necessary for understanding those authors' work. The readers of a scientific paper can instead use the references (or citation list) to appreciate the context of the paper. For instance, looking at the years of publication of each citation a reader can try to guess how obsolete are the contents of the paper.

During the last decade a special use of the citations included in scientific papers is increasing: the evaluation of impact, meaning that if a paper A quotes a paper B this implies that B had some impact on A. Such an evaluation is widely used to measure the quality of a journal and, with ever increasing importance, to decide about the quality of the work of single researchers or teams or even

communities. For instance in Italy in the last few years a new govern agency called ANVUR[3] has set up an evaluation framework for ranking universities and research centres mainly based on the evaluation of the impact of their research products.

Citations are however only a part of the picture. The availability of rich data about scientific papers – including information about authors, publishing processes and, more important, about the content itself – opens the way to novel applications for a large spectrum of users. The same processes of evaluation, access and exploitation of the research results can be improved by combining all this information.

In fact, the knowledge management of scholarly products is an emerging research area: for authors, it includes the gathering of personal repositories of papers, citations, and their relationships with the author's work; for publishers, it includes the construction of large repositories of assets from conferences or journals, sold usually on a subscription base in several digital forms; for universities and research centers, it includes the construction of citational datasets about the scientific papers published by their people; for funding agencies, it includes the ranking of research needs and proposals exploiting ratings based on citational analysis. For all these reasons, there is an ever increasing interest in publishing Linked Open Datasets about scientific papers. However the current landscape is very fragmented: some projects focus on bibliographic data, others on authorship data, others on citations, and so on. The quality is also heterogeneous and the production and maintenance of such datasets is difficult and time-consuming.

In this paper we introduce the Semantic Lancet Project, whose goal is to make available rich semantic data about scholarly publications and to provide users with sophisticated services on top of those data. We developed a chain of tools that produce high-quality data from multiple sources. Such tool chain has been successfully used to produce a rich and freely available LOD, described here as well.

The structure of the paper is as follows: Section 2 describes some related works and the current status of datasets for scholarly publishing; Section 3 delineates some missing pieces and possible improvements for such datasets; Section 4 introduces our Semantic Lancet Project, while Section 5 draws our conclusions.

2 Scholarly publishing and LOD

The idea of creating Linked Open Data (LOD) for scholarly publications is not new. Several repositories of scientific publications have been sided by RDF datasets that can be queried through SPARQL and can be exploited to build sophisticated services for the research community. In this section we give an overview of the most relevant projects and highlight their commonalities, as well as their peculiarities and possible developments.

[3] http://www.anvur.org/

The focus of our analysis is on datasets about scientific articles. There is in fact an ever growing interest in alternative channels for disseminating research results: for instance, publishing raw experimental data [11], using social software platforms and alternative metrics to evaluate research products [10], aggregating papers and supplementary materials in reusable research objects [1], or publishing datasets combining research and educational material [15]. These approaches open challenging and fascinating perspectives but they are left out of the discussion for this work.

The amount and the types of data that could be extracted from scientific articles are boundless. We looked at six orthogonal aspects – *bibliographic data, contributors, citations, affiliations, classification, and abstract.*

The first two are quite obvious: bibliographic data*, i.e.* general information about the papers and their publication (including title, venue, volume, issue, etc.) and data about contributors (authors, editors, publishers, etc.). Such information is published in all datasets, though with very different structures (according to different ontological models) and levels of granularity. These data are often completed by information about affiliations, though there is no dataset we are aware of that describes how that information changed over time. Several datasets also describe citations, given their key role for the research community. There are different ways of making citations available [3] and their current support is very fragmented.

Much interesting information can be extracted and exploited for the final users. The classification of a paper, to identify its type (research, position, editorial, etc.) or to characterize its content (for instance, by using categories of a classification system like ACM or keywords freely assigned by the authors) is also quite common. We think that datasets could be better exploited if the content of the papers was also available as RDF. That would open the way to sophisticated services of content analysis, extraction, recommendation and querying. On the other hand, such information is hard to extract and annotate – either manually or automatically – so that its current support is very low. Thus, we limited our analysis to abstracts only, which are usually freely available, in a few cases, also available as RDF Literals.

The most relevant and complete datasets have been created for the biomedical domain. One of the first was the NPG Linked Data Platform[4]. It includes data about papers published by Nature from 1845 and counts about 400 millions of triples, structured according to Dublin Core, FOAF, PRISM[5] and BiBO[6] vocabularies. Data are very high-quality and cover all aspects we are interested in, though some information is not present for all papers. For instance, only part of the abstracts are available and some papers are classified by genre, others are classified by subject, others are not classified at all. The citation network also is partially covered by the dataset. Two more feature of this dataset are worth highlighting: the fact that it contains data about research products, besides pa-

[4] http://www.nature.com/developers/documentation/linked-data-platform/
[5] http://www.prismstandard.org/
[6] http://http://bibliontology.com

pers, and the fact that is connected to external services, for instance CrossRef[7] to handle citations.

Citations are indeed the key part of the JISC OpenCitation corpus [13, 8], which makes freely available data about papers published in Open Access PubMed Central[8]. The corpus covers about 3.4 million papers and contains a lot of information about them. It primarily uses SPARhttp://purl.org/spar/, PRISM, Dublin Core ontologies. Particularly interesting is the adoption of the PROhttp://purl.org/spar/pro ontology to describe roles and to model authorship information. The dataset also contains several abstracts and data about affiliations, though these are not available for all papers. In general, it is very well-structured and of high-quality but currently not active.

A very similar research is carried on in BioTea [6]. The goal of the project is to make the biomedical literature available as RDF, taking papers again from PubMed Central. The BioTea dataset describes about 270000 papers, published in 2400 journals, according to different ontological models (BiBO, DublinCore, FOAF). Three aspects make this project very relevant to our work. First of all, the automatic workflow for producing the dataset: a set of tools that take as input XML documents and produce rich RDF annotations automatically. Second, the availability of data about the components of the papers. In fact, BioTea relies on the DoCO ontology[9] and contains data about sections, subsections, paragraphs and document objects, that are separate entities in the dataset. Finally, the fact that this work also includes some services built on top of semantic data, for instance a tool for searching human genes in PubMedCentral.

Publishers of computer science papers have also made several datasets available. One of them is DBLP++[10], that makes available RDF data corresponding to those collected in DBLP [7], and coming from multiple publishers and publications. The dataset uses the SWRC ontology[11] [16]. The goal of the original project, i.e. tracking the bibliography of each researcher, does heavily impact the type of information available in the dataset. In fact, it does not contain any data about affiliations nor about the citations. The are instead some abstracts and keywords. The quality of available data is also very fragmented: some papers are fully described, others contain partial data.

Similar issues, even more evident, exist for the dataset representing the publications of ACM[12]. Apart from bibliographic and ACM classification data that are quite complete, the dataset contains only some details of the authors, some citations and a few other information. It is composed of about 12 million triples but it has not been updated since 2006. The description of a lot of papers, in fact, are missing and the dataset is loosely integrated with other sources.

[7] http://www.crossref.org
[8] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[9] DoCO ontology: available at http://purl.org/spar/doco
[10] DBLP++: available at http://dblp.l3s.de/dblp++.php
[11] SWRC ontology: available at http://ontoware.org/swrc/
[12] ACM dataset: available at http://acm.rkbexplorer.com/

A continuously updated dataset is instead the Semantic Web Dog Food[13] [17], containing semantic data about Semantic Web conferences and workshops. Today it counts about 250000 triples on about 5000 papers and 10000 authors. The dataset is explicitly focused on scientific events and bibliographic records about the papers, including those about authors and their affiliations. There are some abstracts and keywords provided by the authors, but there are no data about document components and citations. The corpus is actually composed of different datasets, created and uploaded separately for each conference. That has a great impact on the overall quality, which varies a lot among different conferences. Moreover, it generates a lot of ambiguities and redundant information.

The situation is much easier to control in smaller datasets, associated to single journals. One example is the Semantic Web Journal dataset[14], publishing bibliographic records and rich data of all the papers of the homonymous journal. It contains about 21000 triples, structured according to a specific Semantic Web Journal ontology[15], FOAF, Dublin Core and BiBO. It contains rich bibliographic data, contributions and all the abstracts but a few data about citations. The peculiarity of this dataset, that is actually derived from the peculiar open reviewing process of the journal, is that each paper is also supplied with information about the reviewing process, including intermediate steps, reviewers information, and the reviews.

Other journals are progressively publishing their data as LOD, for instance the Journal of Universal Computer Science[16], that regularly publishes information about the papers, including their bibliographic data, abstracts and topics. Unfortunately citations are still missing in this dataset.

In conclusion, the situation today is fragmented and heterogeneous. None of the datasets we studied covers all aspects, with the same precision and completeness. The integration of multiple datasets is still far from being complete and there are still ambiguities and redundant information that hinder a full exploitation of these sources. In many cases, an automatic process for producing and updating data is missing. Finally, the semantic enhancement of papers' content and the exploitation of such enhancement is still under-explored.

3 Towards better datasets

As seen in the previous section, existing datasets, for one reason or another, are not yet ideal. But what does *ideal* mean, in this context? What are the features that we should expect in new datasets to improve on the current situation? In this section we list a few issues that, in our opinion, should characterize the selection process of data to create better datasets.

Data diversity: Scholarly publishing, of course, exists for all scientific and cultural disciplines, with substantial similarities and analogously relevant differ-

[13] Available at http://data.semanticweb.org/.
[14] Available at http://semantic-web-journal.com:3030/.
[15] http://semantic-web-journal.com/ontology
[16] Available at http://jucs.org:8181/d2rq/.

ences: the types, structures, methodologies and assessments found in literature collections of, say, medicine and archeology are rather different, and shaping the data structure of the LOD along one specific discipline to the detriment of all the others means that such discipline will be most probably the only one to which such LOD can be associated with. A better approach is to identify patterns in literature across disciplines, and possibly collecting as many such patterns as possible, even if applicable only to a few disciplines at all.

Data richness: The richness of the data structure determines the number and expressiveness of possible searches on a dataset. Some data are obvious and undeniably universal: authors, titles, publication venues, and many of the classical bibliographic information sets. Of course there can be much more: is the origin of funding of the research something relevant and worth recording? In many disciplines, this could be considered irrelevant and possibly even non existing. In other disciplines, it is a discriminating factor of the authoritativeness of the research result. What about the overall structure of the paper? Would it be worth to record the existence of a section about related works, about the data collection methodology, about the assessment of the impact of the results? Is the type of the argumentation leading to the main claim relevant to establish the quality of the result? In [19] we proposed a seven level characterization of the types of data that is worth describing of scholarly publication, among which the context (institutions involved, sources of funding, etc.), its structural components (sections, blocks, tabular data, etc.) or its rhetorical structures (introduction, methods, results, etc.).

Data correctness: The richness of the dataset is often the result of heterogeneity in the data sources: each source provides only parts of the information, and the dataset is the union of the data taken from such sources. This implies that the quality of such sources will reflect in the quality of the final dataset, and that harmonization efforts will be required to make sure that information from different sources is coherent. The correctness of the data is therefore an end in itself, as well as the quality and the interconnectedness of the data resulting from the integration of the various sources. This implies both working on the comparison and selection of sources (e.g., when similar data is available from different sources and with individual differences), on the integration of sources (when they complement each other on different aspects of the same entities), and on the actual correction of the data (when the sources, as often happens, contain data of various or dubious quality). Furthermore, such sections could be done automatically (e.g., by the action of rules), or even manually. For instance, the harmonization of the names of publication venues and authors' affiliation, the correct attribution of papers to the right individual (especially in presence of differences in spelling, transliteration, homonyms, etc.), the harmonization of the keywords used to describe the individual works may require a quantity of work both automatic and, if all else fails, manual, that increases the quality of the dataset.

Provenance information: The integration of different data sources with different degrees of correctness, quality, precision and completeness, and the

addition of manual correction of all the errors that have been found, means that the actual origin of each individual piece of information may be due to a number of different actors, may have been the result of a number of different actions, and may have happened in different moments in time. Thus it is important to record everything about the origin and the transformation that each data item has undergone. Cumulatively, this meta-information about the metadata itself is called *provenance* [18]. Especially when a data value has had a troubled and convoluted history, having rich and precise provenance information is crucial to determine its appropriateness and trustworthiness.

Time-awareness: Data changes because things change: our knowledge of things improves and our interpretation of things evolves. For instance, authors affiliation changes (because people move from one employer to another), publications change (because papers are redacted, retracted, extended), significance of experiments change (because new interpretations are provided in subsequent papers), even disciplines change and evolve. Every dataset can be seen either as a frozen description of reality in a given moment, or as an evolving but amnesic organism where every newly available data rewrites and removes previous information, or as an evolving, semi-aware entity that, somehow, remembers and can reflect upon the changes it has undergone. This means being aware of how time and events impact on the data, and being able to reconstruct the dataset as it was at any given moment in time, and to describe the changes that have happened between two moments in time.

Ease of update and enrichment: Obviously, time-awareness and data correctness imply that the content of the dataset evolves and changes regularly, systematically and often profoundly. Allowing easy update and enrichment of the content of the dataset is instrumental in encouraging continuous improvement and correction of the dataset, which is fundamental for its richness.

4 Semantic Lancet Project

The *Semantic Lancet Project*[17] is focused on building a Linked Open Dataset on scholarly publications according to the principles and desiderata listed in the previous sections. Data are freely available and released under the CC0 license[18].

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*[19]. SPAR ontologies are ontological modules that allow one to describe of the various parts of the publishing domain in RDF such as article metadata and citations bibliographic references and citation contexts, person's roles and document statuses, document components and publishing workflows. We suggest to read [20] for a more comprehensive introduction of all the SPAR Ontologies.

[17] The homepage of the project is http://www.semanticlancet.eu.
[18] http://creativecommons.org/about/cc0
[19] Available online at http://purl.org/spar.

On the other hand, we want to make publicly-available a huge and rich RDF triplestore[20] (accompanied by a SPARQL endpoint) and a series of services built upon it of scholarly data starting from those made available from Elsevier's Science Direct[21] and Scopus[22] – while preparing the whole infrastructure in order to facilitate the future managing of data coming from other publishers.

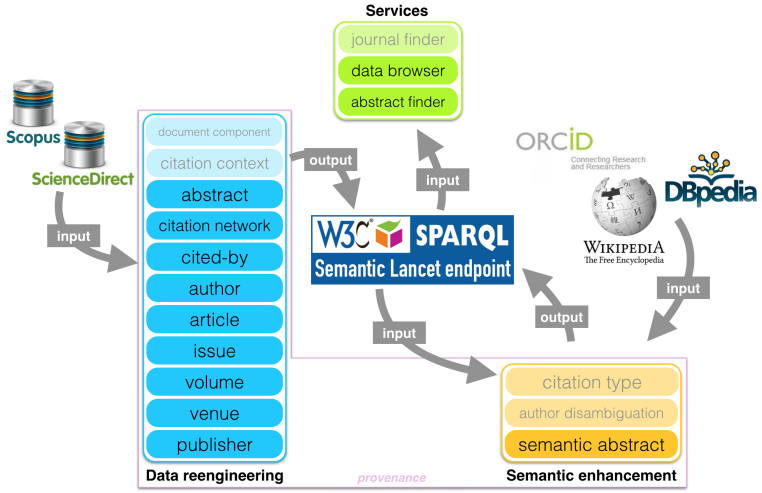

Fig. 1. The overall structure of the three main sections of the Semantic Lancet Project. The blurry blocks (e.g., *document component*, *citation type*, *journal finder*, and the *provenance* module) are currently under development.

The framework of the Semantic Lancet Project, summarised in Fig. 1, is composed basically by three macro sections. First we have the *data reengineering* section that is responsible for the conversion of raw data coming from existing repository (i.e., Scopus and Science Direct) into OWL according to SPAR ontologies and push them to the triplestore. Then, the *semantic enhancement* section uses the data in the Semantic Lancet triplestore and enriches them semantically according to information from different sources. Note that the RDF data result of each of the blocks in the previous two sections are stored in different graphs in the triplestore (usually, one graph for each block). Finally, the *services* section provides applications for browsing and make sense of such data. Each block of a section is actually implemented by a particular script or application and works independently from the others.

In the next section we describe the three main sections of the project with more details. The framework also includes an orthogonal *provenance* module

[20] The triplestore we are currently using is Fuseki. However, we plan to migrate soon to Virtuoso, in order to have more flexibility for handling huge amount of statements.
[21] Science Direct: http://sciencedirect.com.
[22] Scopus: http://www.scopus.com.

(that is currently under development) for adding provenance information about the data added and updated in the triplestore.

4.1 Data reengineering

The data reengineering section is the one responsible of the translation of the raw data coming from the Scopus and Science Direct repositories into RDF. Basically, two kinds of data are requested by using the API made available by Elsevier[23]: those referring to metadata of articles, that are stored in JSON format, and those concerning the full text of articles, stored in XML.

The process of gathering basic article metadata from Elsevier's repositories is handled by a Python script that downloads all the article information available for a given request, on the current date, at both Scopus and Science Direct indexes. 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.

From a preliminary analysis performed considering 10 different journal articles coming from different academic disciplines, we decided to use Science Direct as base repository and Scopus for addressing incorrectness/missing in data.

The process of gathering the full content of articles is handled by an additional script, which communicates with the Science Direct repositories and retrieves the XML file of the full text of articles plus an extended view of the article metadata in JSON as backup, in case the XML is not available.

The collected data are processed by a chain of scripts, shown as separate blocks of the data reengineering section in the left side of Fig. 1. Each script retrieves all the data of interest and convert them into proper SPAR-based RDF statements. Currently the following data are covered:

- all the articles, including their authors, the related venues/volumes/issues, the publisher, and the citation count data as in Scopus *[from the basic metadata of articles]*;
- article abstracts and the whole citation network (in forms of "article A cites article B") of all the articles *[from the full content of articles]*.

We are currently working on additional scripts to extract data about the citation context of citations (i.e., the sentences that contains a pointer to a certain bibliographic reference) and to describe the various components (paragraphs, sections, introduction, related works, methods, materials, evaluation, results, etc.) of each article. Once extracted, all these RDF statements are pushed on the repository through another Python script following the data-model introduced in Fig. 2, and, thus, are made available to Web users for free browsing and download.

[23] Available at http://www.developers.elsevier.com/devcms/content-apis.

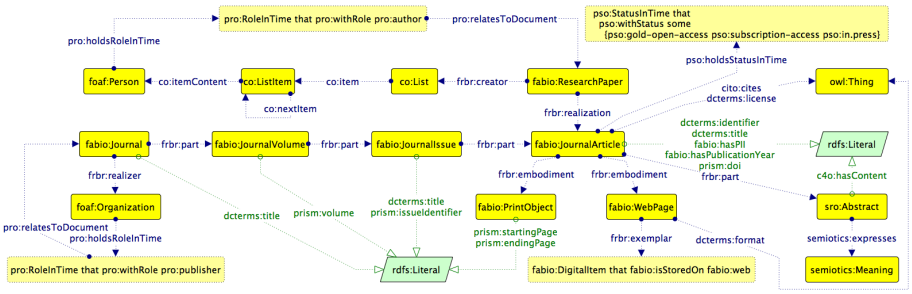

Fig. 2. The data model of all the journal articles in Semantic Lancet.

In order to test the scripts developed, we have already converted and published on the Semantic Lancet triplestore all the data concerning the Journal of Web Semantics[24]. In particular, we have retrieved all the data related to 367 articles, obtaining 80920 RDF statements.

4.2 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. These data are derived from the content of the papers and from external sources; some of them are the result of further refinement of the existing dataset. The more information is made explicit, the more sophisticated functionalities can be provided to the users. The overall framework of Semantic Lancet is in fact modular and we plan to progressively add new tools.

Currently we have implemented a module for generating *semantic abstracts*, which are expressed as RDF/OWL and contain semantic data, extracted from the original abstracts written in natural language text. Note that the original abstracts are available in the dataset, as produced by the *data reengineering* phase.

Such a semantic enhancement typically identifies a variety of tasks aiming at extracting semantic features from unstructured (e.g., natural language text) or semi-structured (e.g., HTML) content and at representing them as RDF. An exhaustive list the tasks and the features extracted can be found in [4] that also provides a comparison among state of the art systems for the semantic enhancement of content.

In this work, we rely on FRED[25][9], which is a tool that implements deep machine reading methods based on Discourse Representation Theory, Linguistic Frames, and Ontology Design Patterns for deriving a logic representation, expressed as OWL, and Linked Data from natural language sentences. FRED

[24] Journal of Web Semantics: http://www.journals.elsevier.com/journal-of-web-semantics/.

[25] http://wit.istc.cnr.it/stlab-tools/fred

also performs named entity recognition and linking, relation finding, taxonomy induction, semantic role labelling, event recognition and word-sense disambiguation. Hence, FRED enables a rich enhancement of abstracts.

For example the sentence below extracted from the abstract of [5] *"The Web Ontology Language (OWL) is a new formal language for representing ontologies in the Semantic Web..."* returns the RDF/OWL representation depicted in Fig. 3 when parsed with FRED.

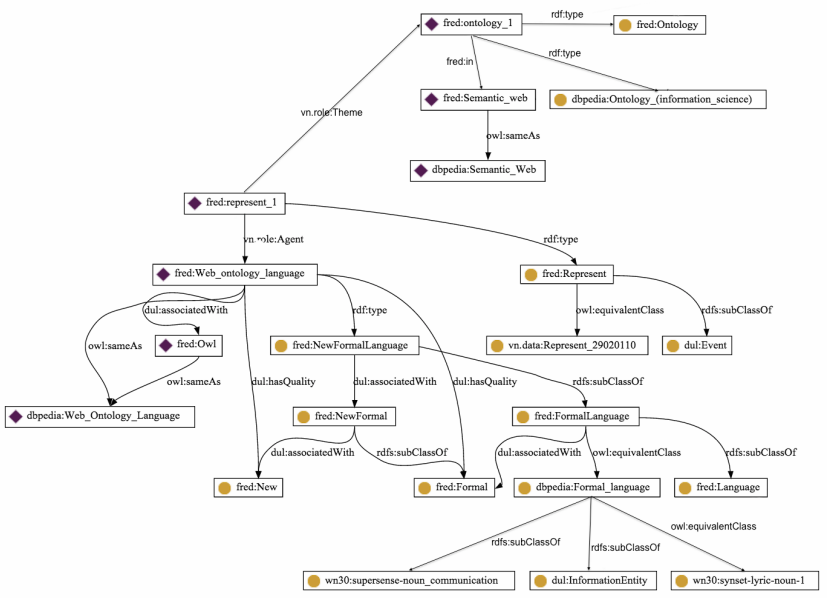

Fig. 3. Semantic enhancement obtained with FRED from the sentence "The Web Ontology Language (OWL) is a new formal language for representing ontologies in the Semantic Web...".

The semantic features extracted from the previous example (cf. Fig. 3) are:

- events, i.e., *fred:Represent*, disambiguated with respect to VerbNet [12];
- semantic roles, i.e., *vn.role:Theme* and *vn.role:Agent*[26];
- named entities, i.e., *fred:Semantic_web*, *fred:Web_ontology_language*,*fred:ontology_1 and fred_owl*, that are also linked when possible to entities in linked data, i.e., DBpedia;
- entity types derived from the natural language text, i.e., *fred:NewFormalLanguage*, with relative taxonomies, i.e., *rdfs:subClassOf* axioms, and detected alignments to WordNet (by means of word-sense disambiguation), D0 and DBpedia, i.e., *owl:equivalentClass* and *rdfs:subClass* axioms.

[26] vn.role:Theme and vn.role:Agent identify the theme and the agent of an event respectively.

All these data can be exploited in novel services for the research community. We have experimented some of these services and briefly describe some other possibilities in the next section.

As shown in Fig. 1, besides the extraction of *semantic abstracts* our approach also encloses additionally and complementary activities, such as, the disambiguation of authors and the citation typing. Hence we are dealing with these activities in our ongoing work. Being able to uniquely recognize an author (aka author disambiguation) is a basic building block for the realization of our vision but, even if named entity resolution and linking (NER) modules of FRED typically produces valuable results, it is not an easy task yet. As a matter of fact, one of the authors of this paper, Andrea Giovanni Nuzzolese, has two entries in the Semantic Web Dog Food (SWDF), namely *swdf:andrea-nuzzolese* and*swdf:andrea-giovanni-nuzzolese* each reporting different facts (affiliation, papers, roles in program committee, etc.) from the same entity but none of the two related to the other by an *owl:sameAs* axiom. This situation results from the fact that NER systems generally take into account only plain literals in which the name of an author appears. Hence, other information such as affiliations, emails, co-authorships, etc. are typically ignored, but de-facto are relevant for author disambiguation.

The citation typing is a further step we are taking. In fact, the frequency a work is cited is a partial indicator of its relevance for a community. More effective results can be obtained by looking for the citation functions, i.e. "the author's reasons for citing a given paper" [14]. Preliminary results have been presented in [2], in which we have introduced CiTalO, a chain of tools for identifying automatically the nature of citations, that we plan to integrate in this project too.

4.3 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 form the others and provides functionalities to the users, built on top of the underlying semantically-enriched data. We have currently implemented two experimental modules, both available in the project website: *data browser* and *abstract finder*.

Data browser. The data browser (cf. Fig. 4) is an interactive and user-friendly interface that allows users to easily access papers, authors, citations, and so on. It is implemented in Bootstrap [27] and compatible with all browsers, including those running in mobile devices. There are two main issues addressed in designing the tool: the huge amount of information users are required to deal with and the complexity of that information. The solution we propose is to hide the intrinsic complexities of the data and of the underlying technologies, giving users an higher-level views over the dataset content. The tool, in fact, does not

[27] http://getbootstrap.com/

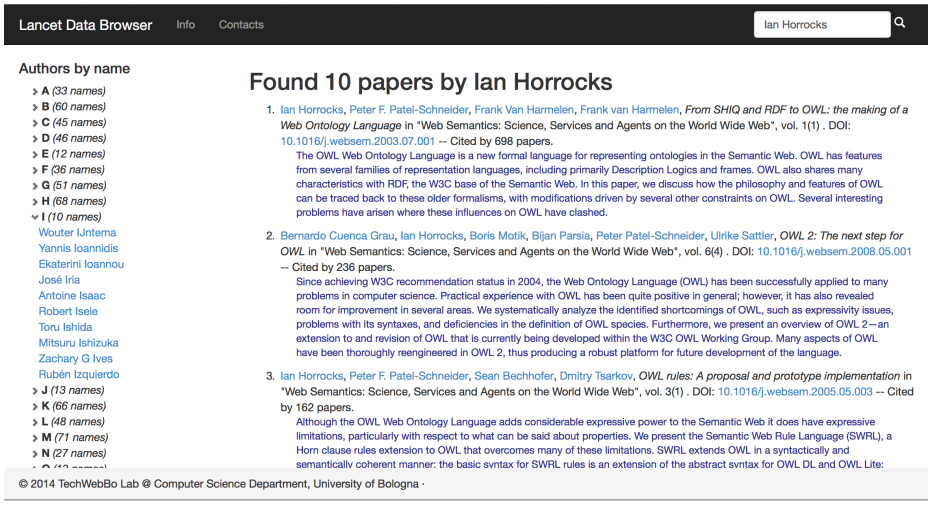

Fig. 4. The Semantic Lancet data browser.

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, and is defined in terms of Work, Expression, Manifestation and Item(s). The dataset items are transparent to the users (though are available to software agents) that only deal with the concept of "paper".

The same happens for semantic properties: there is no distinction between object properties and data properties visible to the users. 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.

Finally, the browser integrates some client-side widgets, like the autocompletion module and the incremental loading of content, speeding up the performances and improving the overall user-experience.

Abstract finder. A further service we are experimenting is for searching *semantic abstracts,* by exploiting the semantic information about concepts, events, roles and named entities produced by FRED. The modules works in two phases: first, it creates a *semiotic index* of the abstracts with respect to a taxonomy of types. These types are aligned to WordNet synsets and DBpedia resources; thus, abstracts are not indexed for their plain textual content but for the concepts represented in that content in a semiotic fashion; second, a simple interface allows users to browse these concepts and the abstracts exposing them;

We plan to extend this module with new functionalities. In particular, we are studying abstract similarity metrics based on the semantic distance among concept maps extracted from textual abstracts (so that, given a paper A and a paper B it will be possible to recognize the similarities or differences between A and B based on their topics) and abstract clustering techniques based on

the analysis of the correlation of topics in the corpus of abstracts and on the application of classification mechanism based on the hybridization of Semantic Web and Machine Learning techniques.

5 Conclusions

There are many other services we can think of built on the Semantic Lancet dataset. One service, for instance, could help authors to find papers relevant for a given research. Towards that goal, we plan to combine both the information in the semantic abstracts – which allow us to identify topics, results and claims of each paper – and the network of citations – which allow us to find the papers considered as related by the authors. Similar services could support not only the authors, but also reviewers and editors in their activities and, as a consequence, provide additional values for the publishers. A machine-readable representation of the papers and, in general, of research activities (and people) could also be exploited for evaluation tasks and help researchers to improve the quality of their scientific production.

Indeed, we need more data to follow this path: authors' affiliations and documents' internal components, just to name a few. Nonetheless, going back to the ideal characteristics of a semantic publishing dataset, we think that the richness of our dataset is acceptable. In fact, we manage to allow users to easily access detailed information about the papers and their abstracts. The integration of multiple sources, cross-checked and merged together, increases the correctness of the dataset. Again, some refinements are still needed, for instance by disambiguating authors' names. The time-awareness issue has been addressed by adopting the SPAR ontologies: the FRBR layered model and the PRO ontology helped modelling the overall publishing process more precisely. Support for the provenance information, on the other hand, is still in progress.

We are daily updating our dataset, which is released in the public domain and open to the community for further experiments and integrations.

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

References

1. Belhajjame, K., Klyne, G., Garijo, D., Corcho, O., García-Cuesta, E., and Palma, R. (2013). Wf4ever Research Object Model. http://wf4ever.github.io/ro/
2. Di Iorio, A., Nuzzolese, A.G., and Peroni, S. (2013). Towards the automatic identification of the nature of citations. In Proceedings of 3rd Workshop on Semantic Publishing (SePublica 2013): 63-74.
3. Di Iorio, A., Nuzzolese, A.G., Peroni, S., Shotton, D., and Vitali, F. (2014). Describing bibliographic references in RDF. Proceedings of 4th Workshop on Semantic Publishing (SePublica 2014). CEUR Workshop Proceedings 1155.
4. Gangemi, A. (2013). A comparison of knowledge extraction tools for the semantic web. The Semantic Web: Semantics and Big Data. Springer, 351-366.

5. Horrocks, I., Patel-Schneider, P. F., and van Harmelen, F. (2003). From SHIQ and RDF to OWL: the making of a Web Ontology Language, Web Semantics: Science, Services and Agents on the World Wide Web, Volume 1, Issue 1, Pages 7-26, ISSN 1570-8268, http://dx.doi.org/10.1016/j.websem.2003.07.001.
6. García-Castro L.J., McLaughlin, C., and García Castro, A. (2013). Biotea: RDFizing PubMed Central in support for the paper as an interface to the Web of Data. J. Biomedical Semantics, 5 (Suppl1):S5.
7. Michael, L. (2002). The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In Proceedings of the 9th International Symposium on String Processing and Information Retrieval (SPIRE 2002), Alberto H. F. Laender and Arlindo L. Oliveira (Eds.). Springer-Verlag, London, UK, UK, 1-10.
8. Peroni, S., Gray, T., Dutton, A., and Shotton, D. (2015). Setting our bibliographic references free: towards open citation data. To appear in Journal of Documentation, 71(2).
9. Presutti, V., Draicchio F., and Gangemi A. (2012). Knowledge extraction based on discourse representation theory and linguistic frames. Knowledge Engineering and Knowledge Management. Springer Berlin Heidelberg, 114-129.
10. Priem, J., Taraborelli, D., Groth, P., and Neylon, C. (2010). Altmetrics: a manifesto.
11. Reichman, O.J., Jones, M.B., and Schildhauer, M.P. (2011). Challenges and Opportunities of Open Data in Ecology, Science 331(6018).
12. Schuler, K.K. (2006). VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon. Unpublished doctoral dissertation , University of Pennsylvania.
13. Shotton, D. (2013). Publishing: Open citations. Nature, 502(7471): 295–297. DOI: 10.1038/502295a
14. Teufel, S., Siddharthan, A., and Tidhar, D. (2009). An annotation scheme for citation function. In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue: 80-87.
15. Zablith, F., Fernandez, M. and Rowe, M. (2012). Production and Consumption of University Linked Data, Interactive Learning Environments, Taylor & Francis.
16. Sure, Y., Bloehdorn, S., Haase, P., Hartmann, J., and Oberle, D. (2005), The SWRC Ontology - Semantic Web for Research Communities. In Proceedings of the 12th Portuguese Conference on Artificial Intelligence - Progress in Artificial Intelligence (EPIA 2005), Springer Berlin Heidelberg, 218-231.
17. Möller, K., Heath, T., Handschuh, S., and Domingue, J. (2007), Recipes for Semantic Web Dog Food - The ESWC2006 and ISWC2006 Metadata Projects. In Proceedings of the 6th International Semantic Web Conference (ISWC2007), Berlin Heidelberg, 802-815.
18. Miles, S. and Gil, Y. (2013). PROV Model Primer. W3C Workin Group Note.
19. Peroni, S., Shotton, D., and Vitali, F. (2012). Faceted documents: describing document characteristics using semantic lenses. In Proceedings of the 2012 ACM symposium on Document Engineering (DocEng 2012): 191-194. New York, New York, USA: ACM. DOI: 10.1145/2361354.2361396
20. Peroni, S. (2014). Semantic Web Technologies and Legal Scholarly Publishing, Law, Governance and Technology Series 15. Cham, Switzerland: Springer. ISBN: 978-3-319-04776-8.