Analysing and Discovering Semantic Relations in Scholarly Data

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

[1] DASPLab, Department of Computer Science and Engineering,
University of Bologna, Italy
{firstname.secondname}@unibo.it
[2] STLab, Institute of Cognitive Science and Technologies,
National Research Council, Italy
andrea.nuzzolese@istc.cnr.it

Abstract. Scholarly publishing has seen an ever increasing interest in Linked Open Data (LOD). However, most of the existing datasets are designed as flat translation of legacy data sources into RDF. Although that is a crucial step to address, a lot of useful information is not expressed in RDF, and humans are still required to infer relevant knowledge by reading and making sense of texts. Examples are the reasons why authors cite other papers, the rhetorical structure of scientific discourse, bibliometric measures, provenance information, and so on. In this paper we introduce the *Semantic Lancet Project*, whose goal is to make available a LOD which includes the formalisation of some useful knowledge hidden within the textual content of papers. We have developed a toolchain for reengineering and enhancing data extracted from some publisher's legacy repositories. Finally, we show how these data are immediately useful to help humans to address relevant tasks, such as data browsing, expert finding, related works finding, and identification of data inconsistencies.

1 Introduction

Scholarly papers are key tools for disseminating, developing and evaluating research results. There is an ever increasing interest in making scholarly data available as Linked Open Data (LOD), on top of which building sophisticated services for the users. The current landscape is variegated, and a lot of information is available through SPARQL end-points and user-friendly interfaces: bibliographic data on journal papers as in BioTea [2], scientific events as in Semantic Web Dog Food [8], citations as in the OpenCitation [18], and so on.

However, existing RDF datasets are built as *conversion of existing data sources into RDF*. A lot of valuable information is still hidden in the text of the papers. Consider for instance the citations: though the bibliographic references to a paper being cited are available, there is no information about the possible reasons why that paper is cited – or at least the context where it was cited. The presence of the abstracts is another example: abstracts are made available as plain text but there is no explicit connection with the entities they refer to, nor a formal representation of their content. Having these data would instead help users to access, understand and compare papers and research results.

We claim that the next generation of LOD of scholarly papers should make such information available, in order to support users (researchers, reviewers, publishers, data curators, etc.) in their daily work and to improve the access to relevant research products. It will be more and more important to provide a broad view of the relations among research papers and to make their content and worth *explicitly formalised* as much as possible.

This paper introduces the *Semantic Lancet Triplestore* (*SLT*), a freely available LOD dataset designed with the aforementioned goal in mind. SLT includes rich data about scholarly papers, that range from a large network of citations (that also includes citation contexts and functions) to semantically-enriched abstracts, from provenance data to time-aware descriptions of the scientific production. The paper also shows how these data can be exploited to complete common tasks in a faster and more effective way. In particular, we show how:

- the presentation of such data can be provided in an intuitive way (supporting users at browsing data about authors and their scientific publications);
- the assessment of the impact of a researcher can benefit from the citation functions (that allow users to have a more precise understanding of the nature citation);
- the search of related works can benefit from semantic abstracts (that allow users to find papers not only with a plain text-based analysis but also taking into account entities, events and roles referred by the abstract);
- the maintenance of such data can benefit from provenance information (that allows users to spot and fix errors and inconsistencies).

SLT has been built with a chain of tools that produces data from legacy sources (i.e., Elsevier's repositories), accessed by REST APIs. The overall workflow – that guarantees a very high level of maintainability and extensibility – is presented in this work as well.

This paper is structured as follows. Section 2 describes some desiderata for a LOD dataset of scholarly papers. Section 3 presents the ontological models used within SLT, and Section 4 describes the reengineering process used to populate the SLT dataset. Section 5 shows how SLT can be exploited to support common research tasks. Section 6 compares the SLT and the current LOD landscape, and Section 7 draws conclusions and future works.

2 Enhancing scholarly papers for analysis and discovery

The idea of creating LOD for scholarly publications is not new. Several repositories of scientific publications have been published as RDF datasets that can be queried through SPARQL and provide information about research results, articles, etc. In this section we discuss some data that LOD triplestores should contain in order to be exploited in sophisticated services for the final users.

The basic building pieces must obviously be the general **(meta)data** about each paper, such as title, venue, volume, issue, citations, and so on. The information about the authors, editors and any other **actor** involved in the publication of each work is also needed, together with data about the **affiliations**. The **abstract** in textual form and the classification of a paper provided by the authors is a further information that we expect. These data are already available in scholarly repositories – for instance, the **subjects** from a taxonomy (as in the case of ACM classification) and the **keywords** – and should be available in RDF datasets as well.

The network of citations is a first area in which a lot of improvement is possible. The minimum requirement is obviously to connect citing and cited papers in a complete network of citations and express it in RDF. Nonetheless, citations are very different in use and scope, because they can have very different functions. It is then important to store them and be able to analyse the original **citation context** [15], i.e., the sentence (or a larger part) of the original paper where a particular work was cited.

The step forward is to make explicit the reason why a paper was cited. As described by Teufel *et al.* [20], the "function of a citation" is the reason why an author cite another paper. This kind of data could be very helpful to understand the nature of each citation and to give it more or less relevance. The automatic identification of the **function of a citation** is not simple – as shown in [1], it is very difficult for humans too and there is a very low agreement in completing such a task – but it can be helpful to better exploit citations.

Many other valuable information can be extracted from the textual content of each paper and made available in RDF datasets. A straightforward application is the extraction of the main **research topics** in a paper, which could be used for automatic reasoning and searching. The issue here is to build a representation that is as faithful as possible to the actual meaning the author had in mind. That is extremely difficult but in some cases it can be encoded through a simple model, e.g., by linking entities to DBpedia in order to represent some semantic aspects of a certain text. In general, the representation as LOD of the content of a paper, from natural language to a graph of entities, enables the development of a new generation of services that can leverage the whole LOD ecosystem for sophisticated access, searching and reasoning.

There are many other scholarly-related data that could be published as LOD. For instance, information about funding agencies or grants associated to each paper. The list of requirements discussed in this section, in fact, is not meant to be exhaustive. However, it allowed us to build both a dataset and a set of applications on top of it that make easier and faster to address some tasks that are specific to the research community, as described in Section 5.

3 Ontologically modelling scholarly knowledge

In this section we briefly introduce the ontologies we used within our framework, providing a bird's-eye view of their main components. Several works have proposed RDFS vocabularies and OWL ontologies for describing particular aspects of the publishing domain, even if they have mainly focused on the description of the metadata concerning bibliographic resources – e.g., DCTerms `http://purl.org/dc/terms`) and BIBO (`http://purl.org/ontology/bibo`)

– rather than their content or contextual information. This is one of the reasons why we built our system around the *Semantic Publishing and Referencing* (*SPAR*) ontologies (`http://www.sparontologies.net`) [12]. SPAR is a suite of orthogonal and complementary OWL 2 ontologies that enable multiple aspects of the publishing process to be described as machine-readable metadata statements, encoded using the Resource Description Framework (RDF).

Particularly interesting for our discussion is the SPAR's ability to enable the characterisation of the nature (or type) of citations, to describe the structure of the paper content and to fully describe agent's roles. The SPAR ontologies in fact allowed us to capture and represent all the aspects that have been introduced in Section 2 – excluding the organisation of the scientific discourse that has been modelled according to the output of one of the tools we have used, i.e. FRED [14] (described in the next section). In particular:

- FaBiO (`http://purl.org/spar/fabio`) and PSO (`http://purl.org/spar/pso`) has allowed us to describe the basic metadata of a publication, as well as its current status (e.g. open-access, subscription-access, in-print);
- PRO (`http://purl.org/spar/pro`) has been used for describing the roles of the contributors of a publication (e.g. author, publisher, etc.);
- BiRO (`http://purl.org/spar/biro`), CiTO (`http://purl.org/spar/cito`) and C4O (`http://purl.org/spar/c4o`) have been used to describe bibliographic reference lists, as well as the citation acts with their related contexts and functions;
- DoCO (`http://purl.org/spar/doco`) has been used to describe the components of a publication (abstracts, sentences, etc.).

4 Building the Semantic Lancet Triplestore

The *Semantic Lancet project*[3] (SLT) is our contribution towards the creation of LOD datasets of scholarly resources, semantically described and enriched as discussed in the previous sections. The aim of the Semantic Lancet Project is twofold. On the one hand, we want to implement a workflow to automatise the production of proper RDF data compliant with the chosen semantic models (i.e. the SPAR Ontologies). On the other hand, we want to make an RDF triplestore of scholarly data publicly-available, starting from the published by Elsevier's Science Direct and Scopus APIs. This is a first step towards the management of heterogeneous data coming from different publishers and repositories.

The three main components of the Semantic Lancet framework are summarised in Fig. 1. First we have the *data reengineering* component that is responsible for the conversion of raw data coming from existing repositories into OWL according to SPAR. Then, the *semantic enhancement* component enriches SLT data semantically according to the information from different sources, such as Wikipedia, DBPedia [6], VerbNet [17], WordNet [7], and Schema.org. Finally, the *provenance* component is responsible for adding provenance information about the data added/modified by the aforementioned two components. In the next subsections we describe these three components in more detail.

[3] `http://www.semanticlancet.eu`

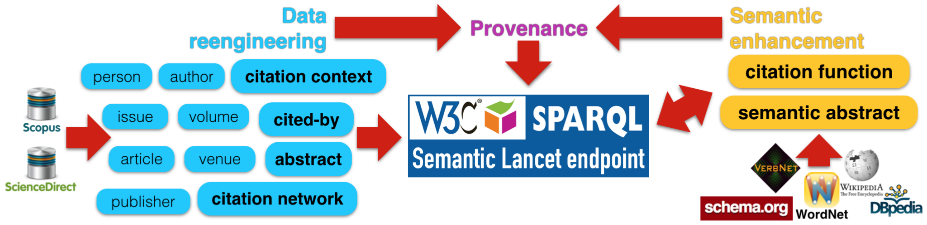


Fig. 1: The Semantic Lancet framework.

4.1 Data reengineering

The data reengineering component (shown in Fig. 1, on the left) takes care of the translation of the raw data coming from Scopus and Science Direct into RDF. Basically, two kinds of data are retrieved using the API made available by Elsevier[4]: the metadata and the full text of the articles. Metadata are retrieved for all papers of a given journal (identified by a ISSN) by querying both Scopus and Science Direct indexes. The full text in XML format is obtained querying Elsevier's repositories by using their Text and Data Mining (TDM) API[5].

These data are used by all the other scripts of our data reengineering process – one script for each blue block of the data reengineering section shown in Fig. 1 – and are converted into proper SPAR-based RDF statements. Finally, all these RDF statements are published on the Semantic Lancet triplestore and, thus, are made available to Web users for free browsing and download.

Achieved results and open issues: during this phase, we had to deal with several issues that emerged from the intrinsic complexity of the data. First of all, the recognition of the sentences was a crucial activity in order to extract citation contexts. Of course, the tokenisation of the sentences has been addressed by means of existing natural language processing tools for sentence recognition. However the main issue we faced was the identification of stop-words that did not mark the end of a sentence despite the fact that they contained "." – for instance "e.g.", "i.e.", "cf.", "a.k.a.", and so on.

Another issue we addressed concerned the creation of the full citation network starting from the XML-based full text of the papers. It was crucial to unambiguously recognise each cited paper, regardless of the bibliographic styles. For this purpose, we are successfully used the Elsevier *Electronic IDentifier* (*EID*) and the *Digital Object Identifier* (*DOI*) as specified in the various bibliographic references of the full text of the downloaded articles, that have to be identified and mined from the actual natural language text fulltext of the papers.

Finally, a side note on a particular issue that is addressed in the next stage of our process, i.e., the disambiguation of people. The approach we currently adopt during this phase is quite basic: we create a new entity in the SLT for each person by normalising the concatenation of the given and family names available in the Elsevier repository. This method allowed us to identify most of the authors but

[4] `http://www.developers.elsevier.com/devcms/content-apis`

[5] `http://www.elsevier.com/about/policies/content-mining-policies`

could not handle correctly composite, incomplete and homonymous names, that are managed in the semantic enhancement stage, as described in Section 4.2.

4.2 Semantic enhancement

The *semantic enhancement* component enriches the triplestore with more semantic data, resulting from a further refinement of those produced during the data reengineering phase, such as abstracts and citation contexts.

Currently, we have implemented a module for generating *semantic abstracts*, i.e., the formal representations of paper abstracts as RDF graphs, defined starting from the original abstracts written in natural language text. To this end, we rely on FRED[6] [14], which is a tool that implements deep machine reading methods based on Discourse Representation Theory, Linguistic Frames and Ontology Design Patterns for deriving a logical representation (expressed in OWL) of natural language sentences. By using FRED, we can also perform named entity recognition and linking to existing entities in the Web and Linked Data (e.g., Wikipedia, DBPedia, VerbNet, WordNet, Schema.org), thus enabling a rich enhancement of textual abstracts. Moreover, other complex tasks on the processed text such as relation finding, taxonomy induction, semantic role labelling, event recognition and word-sense disambiguation are performed. A run of this module queries the triplestore and extracts their related semantic abstracts that will be then linked to the original natural language ones through LMM [13] (i.e., an ontology for expressing semiotical relations) by using the property *semiotics:expresses*.

For example the sentence below extracted from the abstract of [10] *"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. 2 when parsed with FRED.

The semantic features extracted from the previous example (cf. Fig. 2) are:

- events, i.e., *fred:Represent*, disambiguated with respect to VerbNet [17];
- semantic roles, i.e., *vn.role:Theme* and *vn.role:Agent* [7];
- named entities, i.e., *fred:Semantic_Web*, *fred:Web_ontology_language*, *fred:Owl* and *fred:ontology_1*, that are also linked when possible to entities in the 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.

Besides the extraction of semantic abstracts our approach also performs additional and complementary activities, such as the disambiguation of authors. Being able to uniquely recognise an author (aka *author disambiguation*) is a basic building block for the realisation of our vision but, even if FRED typically produces valuable results, it is not an easy task yet. For example, one of the

[6] FRED: `http://wit.istc.cnr.it/stlab-tools/fred`

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

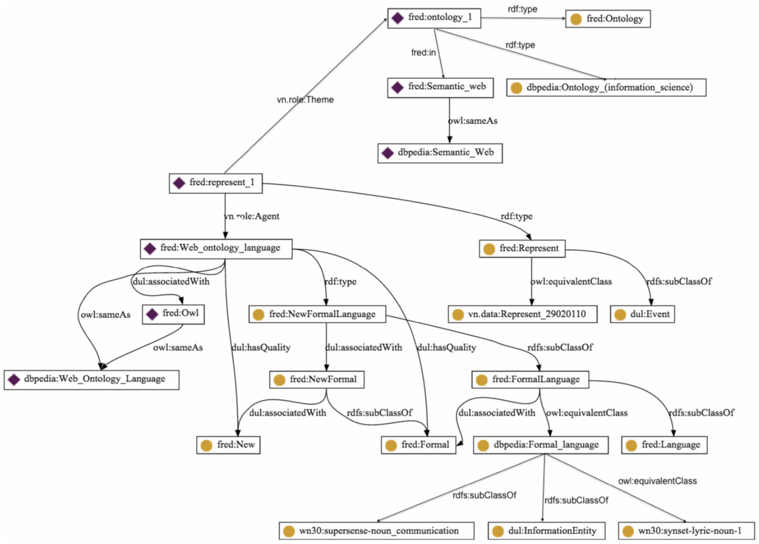


Fig. 2: Semantic enhancement obtained with FRED from the sentence *"The Web Ontology Language (OWL) is a new formal language for representing ontologies..."*.

authors of this paper, Andrea Giovanni Nuzzolese, has two entries in the ScholarlyData dataset [8], namely *sd-person:andrea-nuzzolese* and *sd-person:andrea-giovanni-nuzzolese*. Each entry reports the same facts (affiliation, papers, roles in program committees, etc.) for each 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. To solve this problem we consider other information such as affiliations, emails, co-authorships, etc., which are fundamental to correctly interpret cases such as the aforementioned and successfully perform the author disambiguation task.

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 function*, i.e. "the author's reasons for citing a given paper". For this task, we implemented another module based on *CiTalO* [1], which is a chain of tools to automatically identify the nature of citations according to CiTO (one of the SPAR ontologies), in a way that is comparable with humans. For each paper in the SLT, a run of this module filters all the in-text reference pointers[9] and the related citation sentences introduced during the data reengineering phase, and links them to their related citation functions through OA annotations [16].

[8] The Scholarlydata dataset is the reference linked dataset of the Semantic Web community about papers, people, organisations and events. It is available at `http://www.scholarlydata.org/`

[9] An *in-text reference pointer* is the entity in the body of a citing work that denotes a bibliographic reference in the reference list, e.g. "[3]" and "(Handler et al., 2012)".

Achieved results and open issues: we had to handle several issues for producing data related to semantic abstracts and citation functions, mainly derived from the reliability of the external tools we used.

For the production of the semantic abstracts, one of the problems we addressed was due to the size of the natural language abstracts we used as input. In particular, running FRED with long sentences resulted either in crashes of the service or in quite long computations, which are unacceptable for processing huge quantities of data as in our case. In order to bypass these issues, we split the abstracts into separate sentences (by using a strategy similar to that one introduced in Section 4.1 for citation contexts), ran FRED on each sentence, and merged the results. Even if this process prevented us from using the FRED capabilities for anaphora resolution on multiple and subsequent sentences, it drastically reduced crashes and computational time without a drop in quality.

Another issue was the identification of a citation function in case of CiTalO crashes. Instead of not associating any function to such citation acts, we decided to assign the most generic property in CiTO, i.e., *cito:citesForInformation*, which is the most frequent and neutral citation function (even according to humans, as described in [1]). The main open issue to be addressed still remains how to handle the cases in which these services do not work as expected.

4.3 Data and their provenance information

Since data changes, it is crucial to trace who added and updated each piece of information in the dataset. There is also another tricky issue: data come from different sources. The integration of different data sources with different degrees of correctness, quality, precision and completeness, and the intervention of different agents in the process at different times, means that each piece of information may have been originated by different actors, may have been the result of a number of different actions, or may have been added 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*.

Thus, the SLT actually contains two datasets, one with those data called *Scholarly Data Dataset* (SDD) and one for provenance data called *Provenance Data Dataset* (PDD)[10]. Provenance data are generated by an additional module and stored in the PDD according to the Provenance Ontology (PROV-O)[4], i.e., the W3C Recommendation for tracking provenance information. In particular, we track:

- all the *new RDF statements* created and published in the previous steps (including, for example, their creation date, the agent responsible for such data, the source graph where these data are actually stored in the SDD, and the description of the related creation activities, etc.);
- all the *new provenance data* generated by this step (including temporal information, the description of the creation activities, etc.).

[10] Available at `http://two.eelst.cs.unibo.it/data` and `http://two.eelst.cs.unibo.it/prov`, respectively

These provenance data are helpful to maintain the SDD (as we will discuss in the next section), and are stored separately so as to not interfere with the currently-available scholarly data used by external applications.

5 Exploiting the triplestore

Though SLT can be queried via SPARQL and data are freely available for a smooth integration in the LOD, we want also to show the value of this information for supporting users (e.g., researchers, editors, data curators) in their daily tasks. For this purpose, we have developed a set of tools that provide an interface to the data and enhancement modules developed within the Semantic Lancet project. In particular, in this section we focus on four specific activities: the exploration and analysis of citation networks, the assessment of researchers' relevance in the context of a particular community, the discovery of existing works related to a certain research, and the identification of issues and mistakes in the data published in the triplestore.

5.1 Exploring bibliographies and citation networks

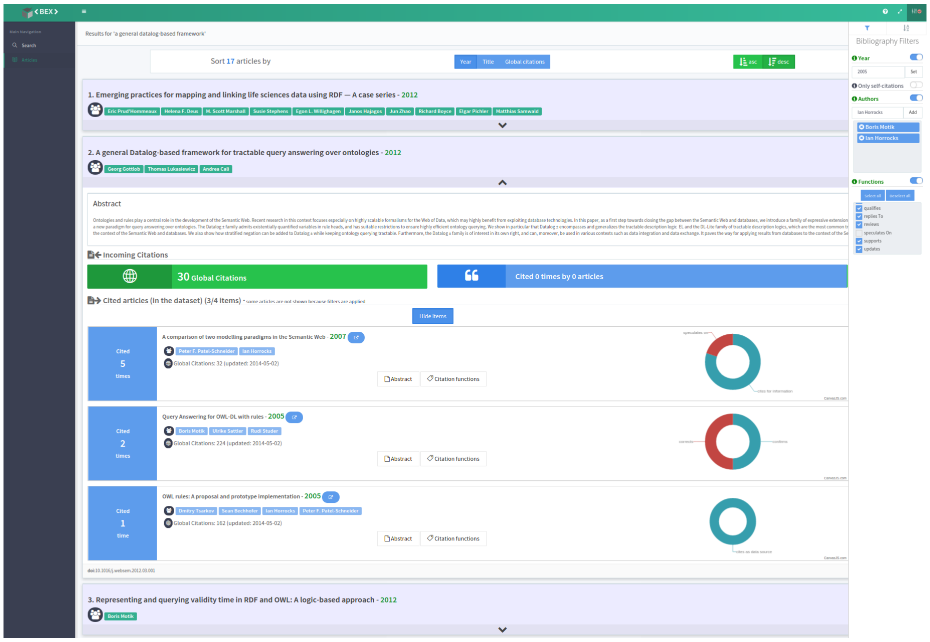
The *Bibliography EXplorer* (BEX)[11][9] is an interactive web-based tool that leverages the rich information about citation networks (i.e. citations functions, citation contexts, etc.) in the SLT to support the analysis, exploration and sense-making process of scientific works. The navigation starts with three search functionalities: besides searching a title or author, the user can also search relevant papers according to their content. This search is performed by calling the Abstract Finder service described in more detail in Section 5.3. Thus, through BEX a user can write in the search box a tentative abstract for her/his paper to retrieve meaningful works that match with it from a pure textual but also semantic point of view. Fig. 3 shows the main interface of BEX and the output of a search.

Search results are organized as a list of papers, ordered by default from the most recent to the oldest one. Through the sorting box at the top of the interface, the user can easily define custom criterion to order the results (i.e. year, number of citations) and the order type (i.e. ascending or descending). For each returned paper, BEX shows a summary of basic information (e.g. title, publication year, author list, etc.) and a link to the official page of the paper on Elsevier's ScienceDirect.

In order to gather more information about a paper in the list, the user can open a sliding box showing the full abstract and data about citations, organized in two separate sections: *outgoing* and *incoming*.

By clicking on the "Show Items" button, the user can get access to the data about the *outgoing citations*, as shown in the central part of Fig. 3. BEX organizes the cited papers in a vertical list. For each cited paper it shows the following information: the number of times in which the paper is referenced by the paper under examination, some general information about the paper, and a

[11] http://eelst.cs.unibo.it:8089/



< BEX > ≡

Main Navigation

Search

Articles

Results for 'a general datalog-based framework'

Sort 17 articles by Year Title Global citations asc desc

1. Emerging practices for mapping and linking life sciences data using RDF — A case series - 2012

Eric Prud'Hommeaux | Helena F. Deus | M. Scott Marshall | Susie Stephens | Egon L. Willighagen | Janos Hajagos | Jun Zhao | Richard Boyce | Elgar Pichler | Matthias Samwald

2. A general Datalog-based framework for tractable query answering over ontologies - 2012

Georg Gottlob | Thomas Lukasiewicz | Andrea Calì

Abstract

Ontologies and rules play a central role in the development of the Semantic Web. Recent research in this context focuses especially on highly scalable formalisms for the Web of Data, which may highly benefit from exploiting database technologies. In this paper, as a first step towards closing the gap between the Semantic Web and databases, we introduce a family of expressive extension
a new paradigm for query answering over ontologies. The Datalog ± family admits existentially quantified variables in rule heads, and has suitable restrictions to ensure highly efficient ontology querying. We show in particular that Datalog ± encompasses and generalizes the tractable description logic EL and the DL-Lite family of tractable description logics, which are the most common tr
the context of the Semantic Web and databases. We also show how stratified negation can be added to Datalog ± while keeping ontology querying tractable. Furthermore, the Datalog ± family is of interest in its own right, and can, moreover, be used in various contexts such as data integration and data exchange. It paves the way for applying results from databases to the context of the Se

Incoming Citations

30 Global Citations

Cited 0 times by 0 articles

Cited articles (in the dataset) (3/4 items) * some articles are not shown because filters are applied

Hide items

Cited 5 times — A comparison of two modelling paradigms in the Semantic Web - 2007
Peter F. Patel-Schneider | Ian Horrocks
Global Citations: 32 (updated: 2014-05-02)
Abstract | Citation functions

Cited 2 times — Query Answering for OWL-DL with rules - 2005
Boris Motik | Ulrike Sattler | Rudi Studer
Global Citations: 224 (updated: 2014-05-02)
Abstract | Citation functions

Cited 1 time — OWL rules: A proposal and prototype implementation - 2005
Dmitry Tsarkov | Sean Bechhofer | Ian Horrocks | Peter F. Patel-Schneider
Global Citations: 162 (updated: 2014-05-02)
Abstract | Citation functions

doi:10.1016/j.websem.2012.03.001

3. Representing and querying validity time in RDF and OWL: A logic-based approach - 2012

Boris Motik

Bibliography Filters

Year: 2005 Set

Only self-citations

Authors: Ian Horrocks Add

Boris Motik

Ian Horrocks

Functions: Select all Deselect all

☑ qualifies
☑ replies To
☑ reviews
☐ speculates On
☑ supports
☑ updates

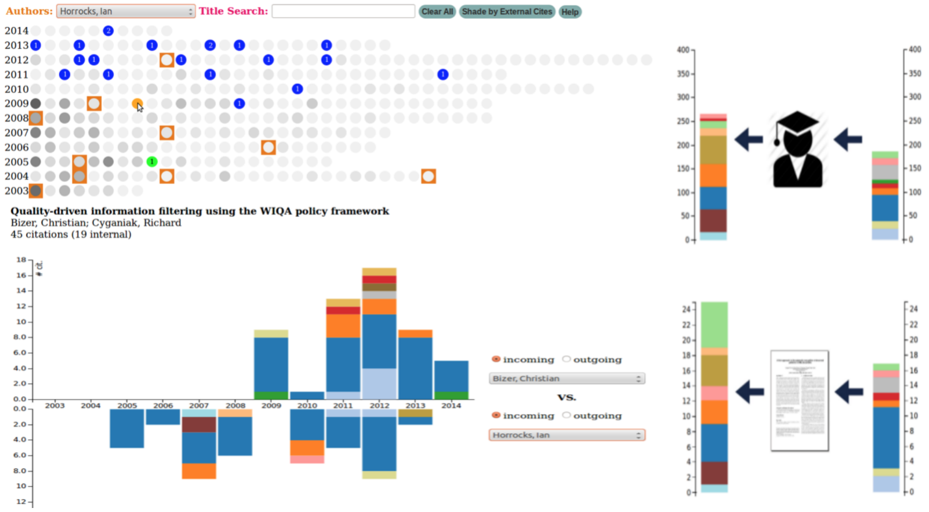
Fig. 3: The main interface of BEX and the list of papers returned by a search. Abstracts and data on citations can be collapsed/expanded on the user's demand.

piechart summarising the number and type of citations received from the focus paper. Moreover, abstracts and citation contexts are shown in popups windows.

In the *incoming citation* section, two counters show the number of *global* and *internal* citations received by the paper under examination. The term 'global' here indicates citations for a paper as counted by external services (Scopus); the term 'internal' indicates the citations given by papers described in our dataset (published in the Journal of Web Semantics).

Further details about the citation functions of incoming internal citations are available. This information is presented in a popup window organized in three parts: a pie chart gives an overview of the number and type of incoming citations (top left), a column chart shows the distribution of the citation functions on a time axis (top right), and details about the citation contexts are presented in the bottom. In the two charts at the top of the page, different colors are used to encode the function of each single citation, and citations with the same function are grouped together. Finally, the last component shows, for each paper citing the paper under focus, the list of the citation contexts.

Finally, BEX provides a rich list of filtering capabilities and ordering criteria that can used to focus on different aspects of the internal citation network. In addition to traditional functionalities, BEX provides additional features, such as filter papers by citation function, and include/exclude the self-citations.



Authors: Horrocks, Ian
Title Search:

Quality-driven information filtering using the WIQA policy framework
Bizer, Christian; Cyganiak, Richard
45 citations (19 internal)

● incoming ○ outgoing
Bizer, Christian
vs.
● incoming ○ outgoing
Horrocks, Ian

Fig. 4: The interface of the Citation Explorer.

5.2 Evaluating the relevance of a researcher

The investigation functionalities provided by BEX, though rich, can be refined by other tools to support other specific tasks that are relevant for scholars' activities and tasks. For example, it is often useful to evaluate the impact and relevance of researchers for a community. Let us consider the following scenario: the editors-in-chief of a journal have decided to give a prize to the most representative and influential author in the history of the journal. The objective of the selection process is evaluating the impact on the community life in terms of both influence and active participation in the debate on the research topics of the journal.

The network of citations can provide valuable information for this task. In particular, a network that also includes some bibliometric indicators and the citation functions provides the editors with a more precise and exhaustive view on the history of each researcher. SLT gives us all the data useful for this analysis, as discussed in Section 2. To help users read and exploit these data, we implemented the *Citation Explorer* [12], an interactive web based tool for analysing and making sense of the citations.

The tool, shown in Fig. 4, is composed by three modules. Once selected a particular journal (i.e., *Web Semantics* in our example), the top-left area shows the overview of all the papers (i.e., the circles) published in the journal and their citation network. This view is based on the attribute-based layout described in [19].

The second module, shown in the right part, summarises the functions (depicted by using different colours) of the incoming and outgoing citations related to each paper (shown in the bottom) and each author (top). This tool provides new elements to evaluate citations: for example, users can grasp the different

[12] http://www.semanticlancet.eu/citationexplorer

impact of an article referenced several times *as an authority* from another cited more times but for a generic reason (e.g., *for information* or *as a related work*).

The last module, shown on the left-bottom, allows users to compare the activity of two authors by showing the distribution of their citations (and the related functions) on a time axis. This time-aware perspective highlights the role played by an author within the particular journal under consideration.

5.3 Searching related works

A third very common task for scholars is searching related works. What a researcher usually does is using a search engine combined with some citation services to make sense of possible interesting articles that talk about a particular topic. However, this is a time-consuming and stressing task to address, since it relies on the ability of the researcher in making connections among papers – that usually must be read totally or in part (e.g., their abstract) to really understand if they are of interest or not. The natural language text of the abstracts (graph *abstract*) and its related formal characterisation (graph *semantic abstract*) available in the Semantic Lancet SDD can be exploited to simplify and, at least, reduce the cognitive effort the researcher uses for addressing such task.

In fact, we have developed a prototypical service called *Abstract Finder* [13]. It is a service for searching relevant papers according to their textual and semantic abstracts, by exploiting the semantic information about concepts, events, roles and named entities produced by the *semantic abstract* module of the semantic enhancement component described in Section 4.2. 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 to the concepts represented in that content. Finally, a simple interface allows users to query for papers having abstracts similar to the input text, and ranks the results according to a similarity measure that takes into account both the textual content of the input text and its formal translation.

Thus, through the Abstract Finder a researcher can, for instance, write in the search box a tentative abstract for her/his paper to retrieve meaningful papers that match with it from a pure textual but also semantic points of view.

5.4 Spotting and fixing data errors

As discussed in Section 4.3 the datasets on scholarly papers need to be maintained and updated. One of the worst nightmares of data curators is to deal with duplicate, incomplete, and inconsistent data. In order to prevent or, at least, to monitor such scenarios, it is valuable to have some mechanisms for debugging datasets and looking for imperfections and errors.

That applies to the SLT as well. In fact, each of the modules in the Semantic Lancet Project pertains the creation of particular kinds of data. What we would

[13] `http://www.semanticlancet.eu/abstractfinder`

like to have, thus, is a dynamic report that spots all the issues of interest, by showing provenance data about them – such as where (i.e., the graph) the problematic RDF statements are stored and what were the modules responsible for their creation. In this way, we can infer whether and when some mistakes have been made, and what we have to fix for addressing such issues.

In order to reach this goal, we have developed the *Web Data Reporter*[14] (WDR), a Web application that queries both SDD and PDD and presents such kinds of situations as a Web page. The current implementation allows us to spot:

- potential *mistakes* in the datasets, e.g., papers that have multiple DOIs;
- data *incompleteness*, e.g., resources associated to no label, type or author;
- data *duplication*, e.g., RDF statements that are defined twice in different graphs.

Finally, since WDR's modular architecture is based on SPARQL queries, new checks and analyses can be easily developed and integrated at any time.

6 Related Work

Several RDF datasets on scientific publications are available today. In this section we summarise the most relevant ones, highlighting their main strengths and weaknesses. The most relevant and complete ones have been created for the biomedical domain. One of the first was the Nature Linked Data platform[15]. It includes data about papers published by Nature from 1845 and counts about 400 millions of triples, structured according to Dublin Core, FOAF, PRISM and BIBO vocabularies. **Pros:** the platform relies on an automatic workflow for converting data into RDF, facilitating its maintainability. It is also connected to external services, for instance CrossRef to handle citations. Data are very high-quality and cover several aspects (e.g. bibliographic metadata, unambiguous authorship information, basic network of citations, content-based article types - such as survey papers, in-use papers, system papers, etc.). **Cons:** however, some information is not present for all papers. For instance, some abstracts are missing, and some papers are classified by subjects, some others by keyword, others are not classified at all. The citation network also is partially covered by the dataset. The citation network is quite basic, with no information about contexts and functions. Moreover, data about authors' affiliations and how they changed over time are missing.

Citations are indeed the key part of the JISC OpenCitation corpus [18], which makes freely available data about papers published in PubMed Central[16]. **Pros:** particularly interesting is the adoption of the PRO ontology to describe roles and to model authorship information, that takes into account time-awareness issues. The dataset also contains several abstracts, data about affiliations ad some classification data, though these are not available for all papers. The website makes available a tool for extracting OpenCitation data from XML sources with an automatic workflow. **Cons:** although the dataset is very well-structured and

[14] `http://www.semanticlancet.eu/reporter`

[15] `http://data.nature.com`

[16] `http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`

of high-quality, one of the main problem is that it is not currently active. Moreover, no provenance data are included, and advanced features about citations (e.g. citation contexts and functions) and article types are missing.

A very similar research is carried on in BioTea [2]. The goal of the project is to make the biomedical literature from PubMed Central 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, Dublin Core, FOAF, etc.). **Pros:** this is a very complete dataset: for example, it contains bibliographic, authorship and affiliation data for all papers. It contains information about the document structure (e.g. sections, subsections, paragraphs, etc.), content fragments are annotated to identify meaningful entities, such as proteins, genes, etc., and the abstracts are fully available. **Cons:** one of the main issue with this dataset is the absence of bibliometric information. Moreover, also some advanced features like data about the rhetorical structure of the papers and citation contexts and functions are missing.

Publishers of computer science papers have also made several datasets available. One of them is DBLP++[17] that makes available RDF data corresponding to those collected in DBLP, and coming from multiple publishers and publications. The dataset uses the SWRC ontology[18]. **Pros:** the dataset contains metadata about papers. Particular attention is given to the keywords and the topics of the papers, on top of which the system provides a facet-based search engine. There are also some abstracts in textual form taken from existing scholarly repositories on the Web. **Cons:** the goal of the original project (i.e. tracking the bibliography of each researcher) does heavily impact the type of information in the dataset. For instance, it does not contain any data about affiliations nor about the citations. The quality of the data is also quite heterogeneous.

The situation is much easier to control in smaller datasets, associated to single journals. For example, the Semantic Web Journal SPARQL end-point[19] publishes bibliographic records and rich data for the homonymous journal. It contains about 21000 triples, structured according to a specific Semantic Web Journal ontology, FOAF, Dublin Core and BIBO. **Pros:** 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 time-aware information about the reviewing process (e.g. reviews, reviewers information, etc.). It contains rich bibliographic data, contributions and all the abstracts. **Cons:** only few data about citations are provided, and information about citation contexts and functions are completely missing. Moreover, no data about affiliations and how they change over the time are provided.

To summarise, the situation today is very fragmented. In particular, the semantic annotation of the content and the exploitation of that enhanced content is still under-explored. SLT plays well in this arena, though it still misses some information that others datasets already make available, even in a larger scale, for instance on affiliations and classifications.

[17] `http://dblp.l3s.de/dblp++.php`

[18] `http://ontoware.org/swrc/`

[19] `http://semantic-web-journal.com:3030/`

7 Discussion and Conclusions

This paper presents a LOD on scholarly papers, a toolchain for producing and updating the dataset, and a set of tools for browsing, investigating and leveraging the information included in the dataset. The triplestore implements our vision of semantic scholarly dataset, that combines basic bibliographic and authorship data with semantically-enriched data extracted from the text of the papers.

Currently SLT contains data about all the papers published in the Elsevier's Web Semantics journal, but we plan to extend it incrementally: the dataset, in fact, is populated and updated by an automatic workflow that generates content from XML sources. Moreover, we are also enriching the set of information included in the dataset. For instance, the next release will include data about authors' affiliations and documents' internal components, just to name a few. The integration of multiple sources, cross-checked and merged together, is a further step: we are also investigating these aspects, experimenting novel interfaces to access SLT and adding support for new tasks. Nonetheless, going back to the ideal characteristics of a semantic publishing dataset, we think that the richness of our dataset is acceptable.

Another key aspect of the project is the overall quality of both the data, applications and services built to support their use. These aspects have been separately evaluated in other previous works, focusing on the different modules composing the SLT project. For what concerns the data reengineering process, all the models used to represent the data in the SLT and discussed in Section 3 have been presented and discussed in our previous work [12]. Also the two main modules that perform the semantic enhancement described in Section 4.1, extracting information from the analysis of natural language texts, have been introduced and evaluated: CiTalO, which computes the citation functions [1], and FRED, the tool used to generate the semantic abstracts [5]. Finally the Bibliography EXplorer (BEX), the main tool presented in Section 5 that provides an interface to the data and the enhancement modules developed for the SLT, has been discussed and evaluated in [9]. As future work, we plan to perform thorough tests on the whole SLT ecosystem for measuring the overall quality of both the dataset and the developed tools.

Acknowledgements This paper was supported by MIUR PRIN 2016 GAUSS Project. We would like to thank Elsevier for granting access to Scopus and ScienceDirect APIs.

References

1. Ciancarini, P., Di Iorio, A., Nuzzolese, A. G., Peroni, S., & Vitali, F. (2014). Evaluating citation functions in CiTO: cognitive issues. In Proceedings of the 11th Extended Semantic Web Conference (ESWC), 580–594. Berlin, Germany: Springer. DOI: 10.1007/978-3-319-07443-6_39
2. García-Castro, L., McLaughlin, C., & García Castro, A. (2013). Biotea: RDFizing PubMed Central in support for the paper as an interface to the Web of Data. Journal of Biomedical Semantics, 5 (Suppl1): S5.

3. IFLA Study Group on the FRBR. (2009). Functional Requirements for Bibliographic Records. http://www.ifla.org/publications/functional-requirements-for-bibliographic-records (last visited November 7, 2016)
4. Lebo, T., Sahoo, S., & McGuinness, D. (2013). PROV-O: The PROV Ontology. W3C Recommendation, 30 April 2013. World Wide Web Consortium. http://www.w3.org/TR/prov-o/ (last visited November 7, 2016)
5. Gangemi, A., Presutti, V., Reforgiato Recupero, D., Nuzzolese, A. G., Draicchio, F., Mongiovì, M. (2016). Semantic web machine reading with FRED. Semantic Web, *Under review.* http://www.semantic-web-journal.net/system/files/swj1297.pdf
6. Lehmann, J., et al. (2015). DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. In Semantic Web., 6(2), 167-195.
7. Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11): 39–41.
8. Muller, K., et al. (2007). Recipes for Semantic Web Dog Food: the ESWC and ISWC metadata projects. In The Semantic Web (pp. 802-815). Springer.
9. Di Iorio, A., Giannella, R., Poggi, F., Peroni, S., Vitali, F. (2015). Exploring Scholarly Papers Through Citations. In Proceedings of the 2015 ACM Symposium on Document Engineering (pp. 107-116). ACM.
10. Horrocks, I., Patel-Schneider, P. F., and van Harmelen, F. (2003). From SHIQ and RDF to OWL: the making of a Web Ontology Language. In Web Semantics: Science, Services and Agents on the World Wide Web, 1(1), 7-26. http://dx.doi.org/10.1016/j.websem.2003.07.001.
11. Ogbuji, C. (2013). SPARQL 1.1 Graph Store HTTP Protocol. W3C Recommendation, 2013. World Wide Web Consortium. http://www.w3.org/TR/sparql11-http-rdf-update/ (last visited December 7, 2016)
12. Peroni, S. (2014). The Semantic Publishing and Referencing Ontologies. In Semantic Web Technologies and Legal Scholarly Publishing, Law, Governance and Technology Series 15: 121–193. Cham, Switzerland: Springer.
13. Picca, D., Gliozzo, A. M., & Gangemi, A. (2008). LMM: an OWL-DL MetaModel to Represent Heterogeneous Lexical Knowledge. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 08).
14. Presutti, V., Draicchio, F., & Gangemi, A. (2012). Knowledge extraction based on discourse representation theory and linguistic frames. In Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2012): 114–129. Berlin, Germany: Springer.
15. Qazvinian, V., & Radev, D. (2010). Identifying Non-explicit Citing Sentences for Citation-based Summarization. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: 555–564. Pennsylvania, USA.
16. Sanderson, R., Ciccarese, P., & Van de Sompel, H. (2013). Designing the W3C open annotation data model. In Proceedings of the 5th Annual ACM Web Science Conference (WebSci13): 366–375. New York, New York, US: ACM Press.
17. Schuler, K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon. http://repository.upenn.edu/dissertations/AAI3179808 (last visited April 1, 2016)
18. Shotton, D. (2013). Publishing: Open citations. Nature, 502(7471): 295–297.
19. Stasko, J. (2014). Value-driven evaluation of visualizations. In Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization (pp. 46-53). ACM.
20. Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 06): 103–110.