Semantic Annotation of Scholarly Documents and Citations

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

[1] Department of Computer Science and Engineering, University of Bologna (Italy)
[2] STLab-ISTC, Consiglio Nazionale delle Ricerche (Italy)
{ciancarini,diiorio,nuzzoles,essepuntato,fabio}@cs.unibo.it

Abstract. Scholarly publishing is in the middle of a revolution based on the use of Web-related technologies as medium of communication. In this paper we describe our ongoing study of semantic publishing and automatic annotation of scholarly documents, presenting several models and tools for the automatic annotation of structural and semantic components of documents. In particular, we focus on citations and their automatic classification obtained by CiTalO, a framework that combines ontology learning techniques with NLP techniques.

Keywords: CiTO, PDF jailbreaking, Semantic Web, citation networks, citation patterns, semantic annotations, semantic publishing

1 Introduction

Researchers can access thousands or even millions of scholarly papers. For instance, currently (June 2013) the ACM Digital Library has 382,000 full text papers and 2,128,000 bibliographic records. There are basically two ways of accessing these sources. First of all, the papers whose identifiers are known (e.g. DOI, or URI) can be downloaded as PDF files, HTML pages, or other formats. Second, most papers include bibliographic records (about authors, editors, publication venue and date, etc.) and lists of references to other works that, in turn, are supplied with similar material. The result is a huge knowledge-base of references, connections between research works, and valuable pieces of content.

We believe that such knowledge bases are still under exploited. One of the reasons is that a lot of interesting information is not available in a machine-readable format. While the human readers can access PDF files and harvest information directly, the software agents can only rely upon a quite limited amount of data for automatic extraction and interpretation or reasoning. Much more sophisticated applications can be built if larger and more expressive semantic information were available.

That is why we are witnessing to an evolution in scholarly publishing driven by the Semantic Publishing community [16] [4] [14]. The idea is to offer a fully-open access to both content and metadata, where rich data on the internal structures of the documents, their components, their (semantic) connections with

other documents and, in particular, their bibliographic references are available as RDF statements and according to appropriate OWL ontologies. This paves the road to the full integration of existing knowledge bases with other Linked Open Data silos, as well as automatic inferences and information extraction.

This paper presents some highlights of our vision and ongoing research on semantic publishing and automatic annotation of scholarly documents, which include models and tools to semantically annotate scholarly documents. In addition, we introduce in more depth the task of automatic characterisation of citations, exploiting *CiTalO*, a tool that takes as input a sentence containing a citation and infers its nature as modelled in *CiTO*[3] [12], an ontology describing the factual and rhetoric functions of citations (part of a larger ecosystem of ontologies for scholarly publishing called *Semantic Publishing and Referencing – SPAR – Ontologies*[4]). Recent developments of CiTalO are presented, with specific attention to the integration with *PDFX*, an online tool that takes a PDF of a scientific article as input and returns an XML linearisation of it.

These tools are meant to be part of a larger platform on top of which novel services for analysing, collecting and accessing scholarly documents will be built. Our long-term goal is to analyse automatically the pertinence of documents to some research areas, to discover research trends, to discover how research results are propagated, to develop new metrics for evaluating research and to build sophisticated recommenders.

The paper is then structured as follows. In Section 2 we introduce some relevant projects in the context of Semantic Publishing and semantic characterisation of scholarly documents. In Section 3 we give an overview of our vision. CiTalO and its recent developments are presented in Section 4. Section 5 concludes the paper presenting some open issues and future works.

2 Related Works

The semantic annotation and enrichment of scholarly documents, and in particular the adoption of Semantic Web technologies, is a hot research topic. In mid-2010, JISC (the Joint Information Systems Committee, a British funding body) funded two sister projects: the *Open Citation project*[5] and the *Open Bibliography project*[6], held respectively by the University of Oxford and the University of Cambridge. Both projects have the broad goal to study the feasibility, advantages, and applications of using RDF datasets and OWL ontologies when describing and publishing bibliographic data and citations.

The Open Citation project aims at creating a semantic infrastructure that describes articles as bibliographic records and their citations to other related works. One of the outcomes of the project was the development of a suite of ontologies, called *Semantic Publishing And Referencing* (*SPAR*), whose aim is

[3] CiTO, the Citation Typing Ontology: http://purl.org/spar/cito.
[4] Semantic Publishing and Referencing Ontologies homepage: http://purl.org/spar.
[5] Open Citation project blog: http://opencitations.wordpress.com.
[6] Open Bibliography project blog: http://openbiblio.net.

to describe bibliographic entities such as books and journal articles, reference citations, the organisation of bibliographic records and references into bibliographies, ordered reference lists and library catalogues, the component parts of documents, and publishing roles, publishing statuses and publishing workflows. SPAR has been used in the Open Citation project as reference model to create a corpus of interlinked bibliographic records[7] obtained converting the whole set of reference lists contained in all the PubMed Central Open Access articles[8] into RDF data. The converted RDF data are published as Linked Open Data.

Similarly, the Open Bibliographic project aimed at publishing a large corpus of bibliographic data as Linked Open Data, starting from four different sources: the Cambridge University Library[9], the British Library[10], the International Union of Crystallography[11] and PubMed[12]. The original publishers' models have been modified to natively include the open publication of bibliographic data as Linked Open Data; furthermore, the scholarly community was continuously engaged in the development of both the ontological model and the final data.

The *Lucero project*[13] is another JISC project, held by the Open University, which aims at exploring the use of Linked Data within the academic domain. In particular, it proposes solutions that could take advantages from the Linked Data to connect educational and research content, so as students and researches could benefit from semantic technologies.

Lucero main aims are:

- to promote the publication as Linked Open Data of bibliographic data[14] through a tool to facilitate the creation and use of semantic data;
- to identify a process in order to integrate the Linked Data publication of bibliographic information as part of the University's workflows;
- to demonstrate the benefits derived from exposing and using educational and research data as Open Linked Data, through the development of applications that improve the access to those data.

The automatic analysis of networks of citations and bibliographic data is gaining importance in the research community. Copestake *et al.* [3] present an infrastructure called *SciBorg* that allows one to automatically extract semantic characterisations of scientific texts. The project was a collaboration between the Computer Laboratory of the University of Cambridge, the Unilever Centre for Molecular Informatics and the Cambridge eScience Centre, and was supported by the Royal Society of Chemistry, Nature Publishing Group and the International

[7] It is available online at http://opencitations.net.
[8] PubMed Central: http://www.ncbi.nlm.nih.gov/pmc/.
[9] Cambridge University Library: http://www.lib.cam.ac.uk.
[10] British Library: http://www.bl.uk.
[11] International Union of Crystallography: http://www.iucr.org.
[12] PubMed: http://www.ncbi.nlm.nih.gov/pubmed/.
[13] Lucero project blog: http://lucero-project.info.
[14] Available at http://data.open.ac.uk.

Union of Crystallography. SciBorg mainly concentrates its extraction efforts on Chemistry texts in particular, but the techniques available were developed to be domain-independent.

Another recent research [10] on these topics has been presented by Motta and Osborne during the demo session of the 11$^{\text{th}}$ International Semantic Web Conference, held in Boston (US). They introduce *Rexplore*, a tool that provides several graphical ways to look at, explore and make sense of bibliographic data, research topics and trends, and so on, which was built upon an algorithm [11] able to identify automatically several relations between research areas.

3 Building applications on annotated scholarly documents

In this section we describe our long term vision, while in the following ones we will go into details of some models and applications we have already implemented. Our overall goal is to study novel *models* for semantic publishing, to design a *platform* that will offer sophisticated services for collecting and accessing scholarly documents, and to develop algorithmic and visual approaches to *make sense* of bibliographic entities and data, as discussed in the following subsections.

3.1 Semantic models for scholarly publishing

We are currently studying how existing formal models for describing scientific documents – such as the SPAR ontologies and the Semantic Web-based EARMARK markup language [8] – can be used to enable the description of scholarly articles according to different semantic facets.

Those facets, we call *semantic lenses* [13], should make it possible to represent document semantics according to different levels of abstraction and to classify all aspects of the publishing production in a more precise, flexible and expressive way (e.g. publication context, authors' affiliations, rhetorical and argumentative organisation of the scientific discourse, citation networks, etc.).

3.2 A platform for academic publishing data

Our aim is to develop a platform to enhance a document enabling the definition of formal representations of its meaning, its automatic discovery, its linking to related articles, providing access to data within the article in actionable form, and allow integration of data between papers. The platform will support authors, publishing houses, and users at large who "consume" a document for any reasonable goal, for instance evaluating its impact on a scientific community.

In particular, the platform aims at helpings developers and common users to address the following tasks:

– evaluating the pertinence of a document to some scientific fields of its contribution;

– discovering research trends and propagation of research findings;
– tracking of research activities, institutions and disciplines;
– evaluating the social acceptability of the scientific production;
– analysing quantitative aspects of the output of researchers;
– evaluating the multi-disciplinarity of the output of scholars;
– measuring positive/negative citations to a particular work;
– designing and including within the platform efficient algorithms to compute metrics indicators;
– helping final users (e.g. researchers) to find related materials to a particular topic and/or article;
– enabling users to annotate documents through interfaces that facilitate the insertion of related semantic data;
– querying (semantic) bibliographic data.

3.3 Making sense of bibliographic entities and data

We are developing tools and algorithms to include in the platform, which aim at understanding and making sense of scholarly documents and their research and publishing contexts.

In particular, we are investigating the best ways to apply well-known Semantic Web technologies – e.g. FRED [15] – other approaches to infer the structural semantics of document components – e.g. the algorithms proposed by Di Iorio *et al.* [7] [6] – to infer automatically the relatedness of different kinds of research, results, and claims, with the intent of automatically linking and recommending set of papers according to a particular topic.

Another work we have done in that direction is the development of *CiTalO*, which we present in detail in Section 4. CiTalO is able to infer the function of citations within a research article, i.e. author's reasons for citing a particular work, and thus can be used to enable the navigation of back and forward citation networks according to a particular dimension, e.g. if the paper is cited positively or negatively.

We believe that the semantically-aware citation networks as well as other semantic characterisations of scholarly papers we are investigating will open new perspectives for evaluating research.

3.4 Scenarios of use as evaluation

The evaluation of the whole work must be performed through task-based user testing sessions. Our main aims are:

– to assess the interaction mechanisms the platform provides to people, e.g. scientists, when accessing the semantic data about the documents so as to re-use or link them with/to other relevant and/or external source of data;
– to understand the usefulness of the platform when it is used for measuring the quality of scientific research within the context of a specific institution, like for instance UniBo;

– to evaluate a recommender system built on the top of the repository of bibliographic entities and data provided by the platform, whose aim is to help final users (e.g. researchers) to look for materials concerning a particular topic and/or article.

As examples, we have identified three use-cases addressing different classes of users in three corresponding scenarios:

– a scenario for an author is to build the "related works" section of a paper under preparation by discovering useful citations. In this case, for verification purposes, we could use a published paper, ignore its related work and bibliography sections, and reconstruct them measuring information retrieval indicators with appropriate metrics;
– a scenario for a reader or a research product evaluator is to start from a document and find documents citing it positively or negatively, and discover in which context;
– a scenario useful for a publishing house or a research institution could be annotating pertinent Wikipedia articles automatically with citations of papers published by the house or institution.

4 Discovering the function of citations

The scenarios presented in the previous section form a roadmap for the next years and involve several interested research groups. The initial development of CiTalO, we presented in past works, and its extension and integration with other external tools for processing scholarly articles stored in different format (such as XML, PDF, LaTeX, etc.), we introduce herein, represents our initial contribution in this direction, which we introduce in the following subsections.

4.1 CiTalO

CiTalO [5] is a tool that infers the function of citations – which are author's reasons for citing a particular paper – by combining techniques of ontology learning from natural language, sentiment-analysis, word-sense disambiguation, and ontology mapping. The tool is available online at `http://wit.istc.cnr.it:8080/tools/citalo` and takes as input a sentence containing a citation and returns a set of CiTO properties characterising that citation[15].

The overall CiTalO schema is shown in Fig. 1. Six steps compose the architecture, that are briefly discussed as follows.

Sentiment-analysis. The first step is optional and consists of running a sentiment-analysis tool on the citation sentence to capture the polarity emerging from the text in which the citation is included. Knowing the polarity of a citation can be eventually exploited to reduce the set of possible target CiTO

[15] As introduced in Section 1, the *CiTO* is an OWL ontology defining forty different kinds of possible citation functions that may happen in a citation act.

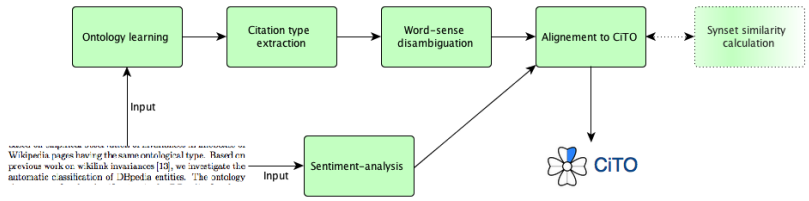

Fig. 1. Pipeline used by CiTalO. The input is the textual context in which the citation appears and the output is a set of properties of the CiTO ontology.

properties, since all properties are clustered around three different polarities: *positive*, *neuter* and *negative*. The current implementation of CiTalO uses AlchemyAPI[16] as sentiment-analysis module.

Ontology learning. The next mandatory step of CiTalO consists of deriving a logical representation of the sentence containing the citation. The ontology extraction is performed by using FRED [15], a tool for ontology learning based on discourse representation theory, frames and ontology design patterns. The output of FRED on a sample sentence "We highly recommend X" is shown in Fig. 2 (also available as RDF triples).

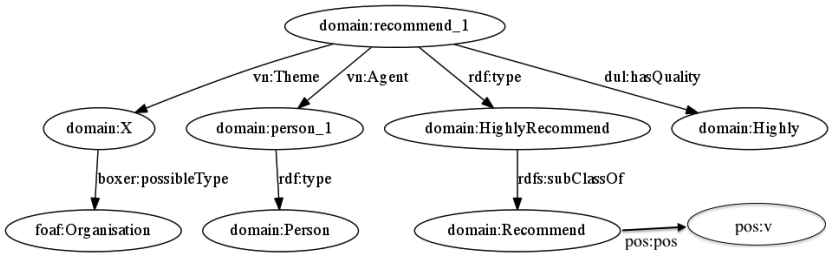

Fig. 2. FRED result for "We highly recommend X".

Citation type extraction. The core step of CiTalO consists of processing the output of FRED and extracting candidate terms which will be exploited for characterising the citation. CiTalO recognises some graph patterns and collects the values of some properties expressed in those patterns. We implemented these operations as SPARQL queries where possible; otherwise, we have directly coded them as Java methods. Considering the aforementioned sentence "We highly recommended X", this step returns the terms *domain:HighlyRecommend* and *domain:Recommend* as candidate types for the citation.

Word-sense disambiguation. The following step consists of disambiguating candidate types and producing a list of synsets that express their sense.

[16] AlchemyAPI: http://www.alchemyapi.com.

To this end, CiTalO uses IMS [17], a word-sense disambiguator, and the disambiguation is performed with respect to OntoWordNet [9]. When running on the last example, IMS provides one disambiguation for the candidate types: *domain:Recommend* disambiguated as *synset:recommend-verb-1*.

Alignment to CiTO. The final step consists of actually assigning CiTO types to citations [12]. We use two ontologies for this purpose: *CiTO2Wordnet*[17] – mapping all the CiTO properties defining citations with the appropriate Wordnet synsets – and *CiTOFunctions*[18] – which classifies each CiTO property according to its factual and positive/neutral/negative rhetorical functions, using the classification proposed by Peroni *et al.* [12]. The final alignment to CiTO is performed through a SPARQL CONSTRUCT query that uses the output of the previous steps, the polarity gathered from the sentiment-analysis phase, OntoWordNet and the two ontologies just introduced. In the case of empty alignments, the CiTO property *citesForInformation* is returned as base case. In the example, the property *citesAsRecommendedReading* is assigned to the citation, derived from *synset:recommend-verb-1*.

Synset similarity calculation. It may happen that none of the CiTO properties can be directly aligned. In that case, CiTalO tries to find a new list of synsets that can be mapped into CiTO properties and are close enough to the ones in the current list. This optional step relies on WS4J (WordNet Similarity for Java) a Java API that measures the semantic proximity between synsets[19].

4.2 Extracting citation functions from PDF

During the *Jailbreak the PDF Hackathon*[20] we attended in Montpellier in May 2013, the main topic under discussion was to enable the extraction of citations out of scientific articles stored as PDF streams. Processing PDF documents is a complex task that requires the use of some sort of heuristics to fully comprehend the actual organisation of document content. The main assumption PDF makes is that any character (and line, picture, and so on) is assigned to a specific location within the page viewport and this declaration can be placed in any position within the PDF stream. Which means that the positions of three different characters making a word such as "who", each following the others according to a sequential order shown in the visualisation of the PDF, can be declared in different and non-sequential places within the PDF stream.

Of course, there exist tools that try to infer and extract the sequential organisation of document content from a PDF stream. One of the most interesting one is *PDFX*[21] [2], which uses *Utopia Documents* [1]. Basically, PDFX takes a PDF of a scientific article as input and returns an XML linearisation of it based on a subset of the *Journal Article Tag Suite* DTD[22], and automatically adds some

[17] CiTO2Wordnet ontology: http://www.essepuntato.it/2013/03/cito2wordnet.
[18] CiTOFunctions: http://www.essepuntato.it/2013/03/cito-functions.
[19] WS4J: http://code.google.com/p/ws4j/
[20] Jailbreak the PDF Hackathon homepage: http://scholrev.org/hackathon/
[21] PDFX homepage: http://pdfx.cs.man.ac.uk/
[22] Journal Article Tag Suite homepage: http://jats.nlm.nih.gov/

structural and rhetorical characterisations of blocks of text according to one of the SPAR ontologies, i.e. the *Document Components Ontology* (*DoCO*)[23].

As a concrete outcome of the hackathon, we started an active collaboration with people (from the Manchester University) responsible for the development of PDFX. From the one hand, they would like to include rhetorical and factual characterisations of citations within the XML document they produce as output of PDFX. On the other hand, we are interested in the development of an automatic process that allows to identify all the citation acts within a scientific document stored as PDF file and then to characterise those citation acts according to the properties in CiTO [12]. We have currently developed such a process and we have also made a Web application called *CiTalO* ^{PDF}, available online[24], to perform the extraction and the semantic characterisation of citations. CiTalO$^{\text{PDF}}$ will be also used by the next release of PDFX so as to add citation functions within the appropriate *xref* elements of the XML conversion obtained by processing the PDF. The workflow followed by CiTalO$^{\text{PDF}}$ is shown in Fig. 3 and defines three different steps, described as follows.

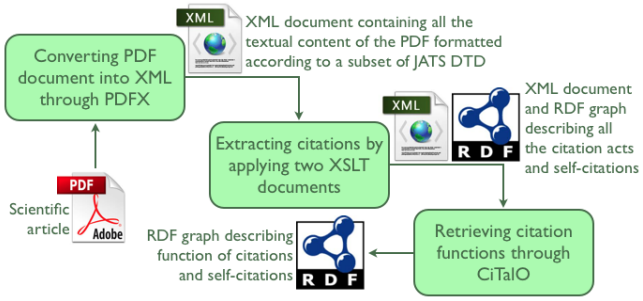

Fig. 3. The workflow followed by CiTalO$^{\text{PDF}}$ to extract citation functions from PDF scientific articles.

Step 1: from PDF to XML. In this step we simply call the PDFX service to convert our input PDF scientific article into an XML document formatted according to a small subset of the JATS DTD.

Step 2: from XML to citation acts. The XML document obtained in the previous step is here processed according to two XSLT documents. The first XSLT document[25] transforms the input XML document into another XML *intermediate* document containing some metadata (element *meta*) of the input document – i.e. the title (element *title*) and the authors (elements *author*) – and all the references (elements *ref*) to bibliographic citations that exist within the document itself, which includes an identifier (element *id*), the label (element *la-*

[23] Document Components Ontology: http://purl.org/spar/doco
[24] CiTalO$^{\text{PDF}}$ is available at http://wit.istc.cnr.it:8080/tools/citalo/pdf/.
[25] http://www.essepuntato.it/2013/citalo/ExtractRefsFromPDFX.xsl

bel) used in the text to point to the bibliographic reference (element *description*) contained in the reference section of the article, and the citation sentence[26] (element *context*) for that citation – where the actual label was substituted by an "X". The following excerpt shows an excerpt of the output of the XSLT process:

```
<refs>
  <meta>
    <title>Anhedonia, emotional numbing, and symptom
        overreporting in male veterans with PTSD</title>
    <authors>
      <author>Todd Kashdan</author>
      <author>Jon Elhai</author>
      <author>Christopher Frueh</author></authors></meta>
  <ref>
    <id>16</id><label>Litz (1992)</label>
    <description id="153">Litz, B. T. (1992). Emotional
        numbing in combat-related post-traumatic stress
        disorder: a critical review and reformulation.
        Clinical Psychology Review, 12, 417-432.
    </description>
    <context>Despite an innovative framework proposed by X,
        it is only recently that emotional numbing and
        anhedonia have received empirical attention in
        studies of post-traumatic stress disorder (PTSD).
    </context></ref> ...
</refs>
```

We apply another XSLT document[27] on the previous XML output, so as to describe the all bibliographic entities involved in citations as an RDF graph. In particular, according to the model illustrated in Fig. 4, we produce an instance belonging to the class *c4o:InTextReferencePointer* (representing an inline pointer to a particular bibliographic reference placed in the references section of the article) for each element *ref* defined in the intermediate document, while the entities describing the scientific article in consideration and the cited articles are all defined as individuals of *fabio:Expression*. In addition, this second XSLT process makes self-citations explicit by adding several statements linking the citing article to a cited one through the property *cito:sharesAuthorsWith* every time at least one of the authors of the former co-authored also the latter.

Step 3: from citation acts to citation functions. In the latter step we actually extend the RDF graph produced before creating an instance of the class *cito:CitationAct* (presented in Fig. 4, representing a citation act) for each element *ref* defined in the intermediate document according to CiTalO's interpretation of the related citation sentence defined in the element *context*. In addition, we explicitly link the inline pointer reference in consideration (defined previously as individual of *c4o:InTextReferencePointer*) to the related citation act by creating an annotation (i.e. an individual of the class *oa:Annotation*).

[26] A *citation sentence* is the sentence where a particular citation act happens.
[27] Available online at http://www.essepuntato.it/2013/citalo/SharesAuthorsWith.xsl.

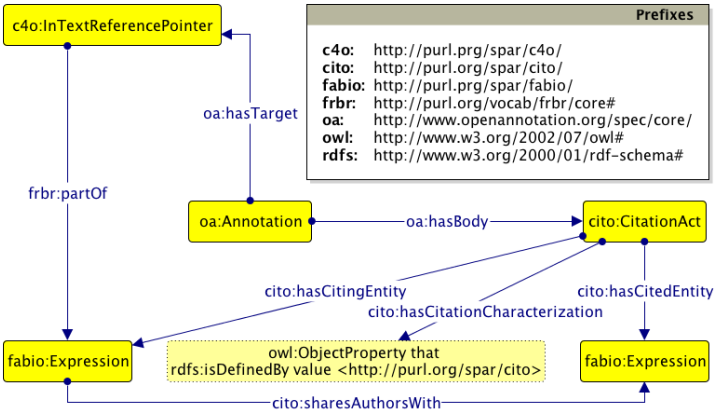

Fig. 4. The diagram defining the model we used to describe the citation acts of scientific documents. Yellow rectangles with solid and dotted border stand for OWL classes (i.e. *owl:Class*) and OWL restrictions (i.e. *owl:Restriction*) respectively, while blue edges define object properties (i.e. *owl:ObjectProperty*) having as domain the class placed under the solid circle and as range the class/restriction indicated by the arrow.

5 Conclusions

In this paper we have described our vision of semantically annotating scholarly papers and some methods and technologies we have developed for this goal.

We are still in an early stage of this research. The availability of large repositories of scholarly documents generates new needs and offers new opportunities. Publishers are building and experimenting new services for the social fruition of scholarly literature. Authors and readers are contributing to establish new social uses. Our results are very initial and still far prom perfect. We need to extend the capabilities of CiTalO when analysing natural language text from scholarly documents, also improving its ability of recognising rhetorical patterns. In addition, we plan to run experiments with several published scholarly articles written in different languages and end users (e.g. researchers).

Acknowledgments We would like to thank Alexandru Constantin and Steve Pettifer for their collaboration and support to the CiTalO framework.

References

1. Attwood, T. K., Kell, D. B., McDermott, P., Marsh, J., Pettifer, S., Thorne, D. (2010). Utopia documents: linking scholarly literature with research data. In Bioinformatics, 26 (18): i568-i574. DOI: 10.1093/bioinformatics/btq383
2. Constantin, A., Pettifer, S., Voronkov, A. (2013). PDFX: fully-automated PDF-to-XML conversion of scientific literature. In Proceedings of the 2013 ACM symposium

on Document Engineering (DocEng 2013): 181-184. New York, New York, US: ACM Press. DOI: 10.1145/2494266.2494271
3. Copestake, A., Corbett, P., Murray-Rust, P., Rupp, C. J., Siddharthan, A., Teufel, S., Waldron, B. (2006). An architecture for language processing for scientific text. In Proceedings of the UK e-Science All Hands Meeting 2006.
4. De Waard, A. (2010). From Proteins to Fairytales: Directions in Semantic Publishing. In IEEE Intelligent Systems, 25 (2): 83-88. DOI: 10.1109/MIS.2010.49
5. Di Iorio, A., Nuzzolese, A., Peroni, S. (2013). Towards the automatic identification of the nature of citations. In Proceedings of 3rd Workshop on Semantic Publishing (SePublica 2013): 63-74. http://ceur-ws.org/Vol-994/paper-06.pdf
6. Di Iorio, A., Peroni, S., Poggi, F., Shotton, D., Vitali, F. (2013). Recognising document components in XML-based academic articles. In Proceedings of the 2013 ACM symposium on Document Engineering (DocEng 2013): 177-180. New York, New York, USA: ACM. DOI: 10.1145/2494266.2494319
7. Di Iorio, A., Peroni, S., Poggi, F., Vitali, F. (2013). Dealing with structural patterns of XML documents. To appear in Journal of the American Society for Information Science and Technology. Hoboken, New Jersey, USA: John Wiley & Sons, Inc.
8. Di Iorio, A., Peroni, S., Vitali, F. (2011). A Semantic Web Approach To Everyday Overlapping Markup. In Journal of the American Society for Information Science and Technology, 62 (9): 1696-1716. DOI: 10.1002/asi.21591
9. Gangemi, A., Navigli, R., Velardi, P. (2003). The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet. In Proceedings of CoopIS/DOA/ODBASE 2003: 820-838. DOI: 10.1007/978-3-540-39964-3_52
10. Motta, E., Osborne, F. (2012). Making Sense of Research with Rexplore. In Proceedings of the ISWC 2012 Posters & Demonstrations Track. http://ceur-ws.org/Vol-914/paper_39.pdf
11. Osborne, F., Motta, E. (2012). Mining Semantic Relations between Research Areas. In Proceedings of the 11th International Semantic Web Conference (ISWC 2012): 410-426. DOI: 10.1007/978-3-642-35176-1_26
12. Peroni, S., Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. In Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 17 (December 2012): 33-43. DOI: 10.1016/j.websem.2012.08.001
13. Peroni, S., Shotton, D., 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. DOI: 10.1145/2361354.2361396
14. Pettifer, S., McDermott, P., Marsh, J., Thorne, D., Villéger, A., Attwood, T. K. (2011). Ceci n'est pas un hamburger: modelling and representing the scholarly article. In Learned Publishing, 24 (3): 207-220. DOI:10.1087/20110309
15. 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 (EKAW2012): 114-129. DOI: 10.1007/978-3-642-33876-2_12
16. Shotton, D. (2009). Semantic Publishing: the coming revolution in scientific journal publishing. Learned Publishing, 22 (2): 85-94. DOI: 10.1087/2009202
17. Zhong, Z., Ng, H. T. (2010). It Makes Sense: A wide-coverage word sense disambiguation system for free text. In Proceedings of the ACL 2010 System Demonstrations: 78-83.