Towards Facilitating Scientific Publishing and Knowledge Exchange Through Linked Data

Sören Auer[1(✉)], Christoph Lange[2], and Timofey Ermilov[3]

[1] Enterprise Information Systems, University of Bonn and Fraunhofer IAIS, Bonn, Germany
auer@cs.uni-bonn.de
[2] School of Computer Science, University of Birmingham, Birmingham, UK
math.semantic.web@gmail.com
[3] AKSW Research Group, University of Leipzig, Leipzig, Germany
ermilov@informatik.uni-leipzig.de

Abstract. In this position paper, we describe our vision of an architecture of participation for semantic linking and contextualizing of research articles. We discuss requirements of such an architecture and showcase an early first prototype.

The Linked Data paradigm has recently evolved into a powerful enabler for integrating structured information and data on the Web and within Enterprise intranets. It is based on the RDF data model and de-referenceable URIs, which not only allows for describing resources and linking to them, but also accessing them using the HTTP protocol to retrieve structured information.

Scientific knowledge exchange (cf. Fig. 1) often involves structured information, such as experimental results, collected data, taxonomies or formulas. Data portals can be used to publish data underlying a certain publication. However, even the actual text of scientific publications often contains structured information currently hidden in prose. Examples include (a) claims and supporting evidence for these, (b) related approaches with their advantages and disadvantages, or (c) a taxonomical classification of the approach described in a certain publication. Such information could easily be expressed and represented in a structured way in RDF. Once scientific publications are increasingly represented in a way that preserves the structure of information, related or similar information from different publications can easily be interlinked and integrated. A survey on a certain research area, for example, could then possibly be generated almost automatically, by collecting the taxonomic classifications as well as advantages and disadvantages of various approaches from different structured publications. As a result, scientific knowledge sharing would be improved substantially, since researchers and other stakeholders would be enabled to search and discover research results not only by using keyword search and following citations, but by formulating sophisticated queries such as "List me all Named Entity Recognition approaches published in the last 5 years, together with the corresponding precision and recall they achieve on a certain benchmark corpus". Currently, answering such a

Ł. Bolikowski et al. (Eds.): TPDL 2013, CCIS 416, pp. 10–15, 2014.
DOI: 10.1007/978-3-319-08425-1_2,

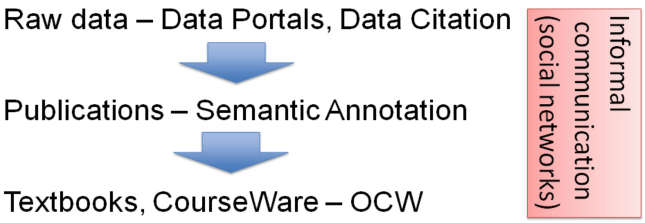

Fig. 1. Different means of scholarly communication.

relatively simple question costs a researcher several weeks or even months of research. Especially for young researchers its currently extremely difficult to navigate through the jungle of research related to their research question.

Although a general solution for this problem is relatively straightforward to realize – researchers could simply publish some RDF Linked Data describing their research along with a paper – the main challenge is to create a network effect through an architecture of participation. This is required, since very few researchers would spend the additional effort of creating a semantic description in addition to a paper if the benefit of doing so would not be immediate. We discuss some requirements, challenges and possible solutions for realizing this vision of truly semantic scientific knowledge exchange.

1 Requirements for Linking and Contextualizing Research Articles

With increasing provision of linked data vocabularies for representing knowledge in specific fields of science[1] but also across science[2], we no longer see the bottleneck on the side of representing and publishing scientific papers as linked open data, but on the side of knowledge acquisition from readers and authors.

For obtaining a critical mass of Linked Data from research papers, we are interested in an approach that is practically feasible, that attracts a large number of users, and that poses a low entry barrier to them.

Practical feasibility means that we do not currently expect a strong natural language processing (NLP) algorithm to fully automatically extract a sufficient RDF graph from

[1] For mathematics and all sciences involving mathematical formulas, see, for example, [1]. Pointers to vocabularies for further scientific domains, particularly including biology and medicine, a long-standing stronghold of semantic web applications, can be found in the Linked Open Vocabularies dataset (http://lov.okfn.org/dataset/lov/).

[2] See, for example, the BIBO bibliographical ontology (http://bibliontology.com/), and the SALT ontologies for rhetorical structures and claims [2], and the more recent SPAR family of Semantic Publishing and Referencing ontologies (http://purl.org/spar/), whose Document Components Ontology DoCO reuses the SALT Rhetorical Ontology and whose FaBIO ontology is more expressive while at the same time computationally more efficient than the still widely used BIBO (cf. [3]).

a paper's full text. Instead we rely on the partial application of well-tried NLP techniques such as named entity recognition, but primarily expect users to manually complete the annotation of the paper, supported by an assistive user interface.

Attracting a large number of users means that we have to work with the most widely used document format for scientific publications, which at this point is PDF, and that our target audience should comprise all readers of publications rather than just their writers. From a knowledge acquisition point of view it may be of advantage to tap the author's stream of consciousness by an invasive editing approach, where semantic annotation facilities seamlessly integrate into the author's preferred editor. Invasive editing solutions (cf. the Related Work section below) promise to reduce the author's effort of inserting frequently occurring structures into the document, while at the same time capturing the precise semantics of these structures. However, solutions would have to be as diverse as the editors that authors prefer, and invasive editing does not cover semantics that not the author but the reader of a publication may think of, e.g. related work. Readers rarely have access to the authors' source documents but rather just to PDFs created from them. Where publishers ask for the sources, which most commercial publishers do, they use them internally, e.g. for typesetting, but do not usually make them available. Some open access publishers, such as *arXiv*, publish sources, whereas most sites for sharing publications, e.g. *ResearchGate* or *SlideShare*, do not support all common source formats;[3] as a result, most users upload PDF. Thanks to the wide adoption of Adobe Reader there is not such a diversity of PDF readers as of document editors; however, Adobe's dominance is decreasing, with recent versions of the Chrome and Firefox web browsers providing their own integrated PDF readers and the default PDF readers of Mac OS X and common Unix desktop environments catching up with Adobe Reader in functionality.

Posing a low entry barrier to users means that the user interface for annotation should "invade" the reader's preferred reading interface as seamlessly as possible. Providing an annotation plugin for a PDF reader is more challenging than developing an annotation plugin for a web browser. Adobe Reader offers scripting support similar to web browsers[4], but the problem is that the PDF format is designed for layouting pages. Well-behaved authoring tools can be configured to preserve some of the original structure of a text, e.g. words before hyphenation, but authors and publishers still rarely pay attention to such aspects. However, it is inherently impossible to fully preserve the original text in a PDF. When a paragraph crosses a page break, selecting that paragraph will always include the footer of the first page and the header of the second page, thus making it impossible to precisely annotate the paragraph.

We lack an architecture of participation for linking and contextualizing research articles. In order to realize such an architecture, we need to provide instant benefits for semantic annotations (e.g.: find related work, gain reputation on social networks, visualization, fun) as well as medium and long-term benefits for semantic annotations (e.g. being cited by more authors, or being more visible to funding bodies).

---

[3] ResearchGate and SlideShare only allow single-file uploads, which is suitable for office documents but hardly for LaTeX documents, which usually involve multiple source files.

[4] http://www.adobe.com/devnet/acrobat/javascript.html

2 Example and Prototype

The following example code in RDF/Turtle shows a possible annotation for a paper describing a novel link discovery approach as well as its implementation and evaluation.

```
limes-paper  describes   appr123 , impl123 , eval123 .
appr123      a           Approach ;
             for         Link_Discovery ;
             hasProp     lossless .

impl123      a           Implementation ;
             implements  appr123 ;
             language    Java .

eval123      a           Evaluation ;
             evaluates   impl123 ;
             uses        DBpedia .
```

Figure 2 shows the early prototype of a semantic annotation platform[5], where an article is shown on the left hand side and an annotation panel on the right. When a user selects a certain part of the article (e.g. a named entity, paragraph, table etc.) an annotation can be added on the right, describing what the selected element represents as well as its features. For example, the section describing the implementation can be annotated with features, such as the programming language chosen for the implementation. During the process of adding annotations, existing properties, concepts and entities are suggested to the user for reuse. As a result, annotations are not isolated but reuse existing vocabulary and establish semantic links between annotated papers. An instant benefit for the user is then, for example, the retrieval of similar articles as shown in the lower right corner of Fig. 2.

3 Related Work

Invasive editing in traditional authoring software has, for example, been realised for mathematical and rhetorical structures of knowledge, by semantic macro packages for *LaTeX* [2, 4], and by plugins for PowerPoint [5]. None of these solutions has been adopted widely. Of the three examples given, only sTeX is still being maintained. With HTML5 advancing, lightweight invasive editing solutions have more recently been realised in web interfaces, which have been extended to enrich the HTML document being authored with RDFa annotations. Examples include the RDFa Content Editor RDFaCE [6] and the One Click Annotator [7]. Both are based on TinyMCE, an HTML editing component widely used in web content management systems. A similar JavaScript-based architecture could be adopted by a browser plugin for annotating read-only HTML documents published on the Web.

[5] The prototype is based on the PDF.js plugin bundled with recent Firefox browsers (https://github.com/mozilla/pdf.js). Source code is available at https://github.com/AKSW/semann.

Fig. 2. Prototype of a semantic annotation platform, with document display (left) annotation panel (upper right) and semantic similarity search (lower right).

4 Conclusions

Exploring new ways of how scientific knowledge can be shared is a very promising area of research and technology. While a number of approaches for semantic annotations and representations of scholarly content exist, an architecture of participation, where researchers are instantly gratified for contributions in the form of small semantic annotations created while reading, is still missing. In this article we presented some requirements as well as an example and first prototype for a semantic annotation platform for research articles.

References

1. Lange, C.: Ontologies and languages for representing mathematical knowledge on the semantic web. Semant. Web J. **4**(2), 119–158 (2013). http://www.semantic-web-journal.net/content/ontologies-and-languages-representing-mathematical-knowledge-semantic-web
2. Groza, T., Möller, K., Handschuh, S., Trif, D., Decker, S.: SALT: weaving the claim web. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K-Il, Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 197–210. Springer, Heidelberg (2007)
3. Shotton, D., Peroni, S.: Libraries and linked data #5: using the SPAR ontologies to publish bibliographic records. Semantic Publishing Weblog (2013). http://semanticpublishing.wordpress.com/2013/03/01/lld5-using-spar-ontologies/
4. Kohlhase, A., Kohlhase, M., Lange, C.: sTeX – a system for flexible formalization of linked data. In: I-Semantics. ACM (2010). http://kwarc.info/kohlhase/papers/isem10.pdf
5. Kohlhase, A.: Semantic interaction design: composing knowledge with CPoint. Ph.D. thesis, University of Bremen (2008)
6. Khalili, A., Auer, S., Hladky, D.: The RDFa Content Editor – From WYSIWYG to WYSIWYM. In: Proceedings of COMPSAC 2012 – Trustworthy Software Systems for the Digital Society. IEEE (2012). http://svn.aksw.org/papers/2012/COMPSAC2012_RDFaCE/public.pdf
7. Heese, R., Luczak-Rösch, M., Oldakowski, R., Streibel, O., Paschke, A.: One click annotation. In: Proceedings of the Sixth Workshop on Scripting and Development for the Semantic Web (SFSW), CEUR-WS.org Workshop, vol. 699 (2010). http://ceur-ws.org/Vol-699/Paper4.pdf

http://www.springer.com/978-3-319-08424-4

Theory and Practice of Digital Libraries -- TPDL 2013 Selected Workshops
LCPD 2013, SUEDL 2013, DataCur 2013, Held in Valletta, Malta, September 22-26, 2013. Revised Selected Papers
Bolikowski, Ł.; Casarosa, V.; Goodale, P.; Houssos, N.; Manghi, P.; Schirrwagen, J. (Eds.)
2014, XV, 250 p. 72 illus., Softcover
ISBN: 978-3-319-08424-4