Proceedings of the First Workshop on Semantic Publication (SePublica2011)

8th Extended Semantic Web Conference Hersonissos, Crete, Greece, May 31, 2011

edited by Anita De Waard, Alexander García Castro, Christoph Lange, and Evan Sandhaus

May 31, 2011

Preface

This volume contains the papers presented at SePublica2011 (`http://sepublica.mywikipaper.org`): First International Workshop on Semantic Publications held on May 30, 2011 in Hersonissos, Crete, Greece.

There were 7 submissions. Each submission was reviewed by three program committee members.

We would like to thank our sponsor Elsevier for generously funding a best paper award, as well as our peer reviewers for carefully reviewing the submissions and giving constructive feedback.

This proceedings volume has been generated with EasyChair, which made this task really convenient.

May 5, 2011
Bremen

Anita De Waard
Alexander García Castro
Christoph Lange
Evan Sandhaus

Program Committee

Christopher Jo Baker	University of New Brunswick
Paolo Ciccarese	Harvard Medical School & Massachusetts General Hospital
Tim Clark	Massachusetts General Hospital / Harvard Medical School
Oscar Corcho	Universidad Politécnica de Madrid
Stéphane Corlosquet	
Joseph Corneli	Knowledge Media Institute, The Open University
Anita De-Waard	Utrecht University
Michael Dreusicke	
Henrik Eriksson	
Alexander Garcia	Postdoctoral fellow
Leyla Jael García Castro	Universitaet der Bundeswehr
Benjamin Good	The Genomics Institute of the Novartis Research Foundation
Tudor Groza	
Michael Kohlhase	KWARC
Sebastian Kruk	
Thomas Kurz	Salzburg Research Forschungsgesellschaft
Christoph Lange	Jacobs University Bremen
Steve Pettifer	The University of Manchester
Matthias Samwald	Digital Enterprise Research Institute, National University of Ireland Galway, Ireland
Evan Sandhaus	New York Times
Jodi Schneider	DERI, NUI Galway
Dagobert Soergel	College of Information Studies, University of Maryland

Additional Reviewers

H

Hindle, Matthew M.

Contents

Authoring and Publishing Units and Quantities in Semantic Documents

Mihai Cîrlănaru, Deyan Ginev, Christoph Lange

Computer Science, Jacobs University Bremen, Germany
{m.cirlanaru,d.ginev,ch.lange}@jacobs-university.de

Abstract. This paper shows how an explicit representation of units and quantities can improve the experience of semantically published documents, and provides a first authoring method in this respect. To exemplify the potential and practical advantages of encoding explicit semantics regarding units w.r.t. user experience, we demonstrate a *unit system preference* service, which enables the user to choose the system of units for the displayed paper. By semantically publishing units, we obtain a basis for a wide range of applications and services such as *unknown unit lookup*, *unit and quantity semantic search* and *unit and quantity manipulation*. Enabling semantic publishing for units is also presented in the context of a large collection of legacy scientific documents (the arXMLiv corpus), where our approach allows to non-invasively enrich legacy publications.

1 Motivation

Units and quantities, although widely spread, lack a formal standard representation for semantic publishing. A multitude of problems [Usm] arise from the different flavors (country specific unit standards) and formats (abbreviations, special cases of occurrence) of units, making it hard for the untrained reader to fully understand the information provided. Semantic publishing solves most such problems by disambiguating the unit and quantity occurrences, which, further on, will enable a wide range of applications and services to interact with them.

A **unit** is *any determinate quantity, dimension, or magnitude adopted as a basis or standard of measurement for other quantities of the same kind and in terms of which their magnitude is calculated or expressed* [Oxf], but from the top-most level of perception, it simply provides information on a wide range of quantifiable aspects. Concrete examples for the great extent of units and quantities include cooking recipes, medical prescriptions, scientific papers and many other. Semantic publishing can provide the middle layer that would ensure a (automated) way of identifying and understanding these occurrences which can enable the evolution of useful technologies and services.

At the perception level, aside from quantifying properties and relations between objects, units bring the meaning of scale. Moreover, units have allowed scientists to better transmit and exchange knowledge among themselves.

In real life, the misinterpretation of units and their quantities has often caused accidents with harsh/expensive consequences. Consider losing a $125 million satellite [Mar] because of the differences between metric and imperial unit systems, or running out of fuel in mid-flight with an aircraft whose fuel sensors were faultily configured in displaying the units [Air]. Fields like medicine, commerce, civil engineering have also been marked by such types of errors and pitfalls [Usm]. This simply emphasizes the fact that units are frequently misinterpreted.

Providing semantics to units and their quantities for the publishing industry, either by supplying semantic authoring tools or by semantically enriching their occurrences in legacy documents, has high impact benefits. It will enable transparent exchange of scientific knowledge between different academic communities, typical of technical papers with high occurrence of units and quantities, and also enhance the reader's experience, via novel interactive services with day-to-day published material, e.g. cooking recipes or technical manuals.

In the following sections the preliminaries (section 2) and state of the art (section 4) for *units and quantities* are introduced in order to have a basis for the *unit and quantity interaction services* (section 6) presented in this paper. We outline immediate strategies (section 5) for extending the benefits of semantic units to legacy documents (section 7) and conclude with a summary of our mid-term outlook of future work (section 8).

2 Preliminaries

The core of semantic publishing resides in open and standardized markup languages used to encapsulate semantics. OpenMath and *Content* MathML are the most widely used semantic markup (also called "content markup") languages for mathematical expressions, which are ubiquitous in science and engineering.

2.1 OpenMath and Content MathML

OpenMath [Bus+04] and the semantically equivalent Content MathML [Aus+10] are standards for the representing the semantics of mathematical expressions [KR09] – as annotations to visual renderings, or for the purpose of communication between computational services. Our investigations focus on these two languages.[1]

Structurally, both OpenMath and MathML provide a valuable basis for machine processing of mathematical expressions; they are ideal markup languages for the purpose of semantic publishing of units and quantities. The expressivity of MathML, provided by its vocabulary having close to 100 XML elements for

[1] The prevalence of XML-based semantic markup languages for representing mathematical expressions – as opposed to RDF – has historical reasons but is also due to the complex n-ary and ordered structures of mathematical expressions, which are hard to break down into RDF triples. In both representations the vocabulary terms (here: functions, operators, sets, constants) are identified by URIs. We refer to [Lan11] for an in-depth treatment.

functions and operators for mathematics [KR09] and multiple *unit and quantity* representation possibilities [DN03], and the modularity and extensibility of OpenMath's vocabulary by way of modular ontologies ("Content Dictionaries", abbreviated as CDs), enable the development of applications and services (some of which are discussed in section 6.2) that build upon the semantic publishing of units and quantities.

2.2 The Semantic Publishing Pipeline

Semantic Publishing, as a process, consists of at least three components, namely *authoring*, *publishing* and *interaction*. Usually these processes imply three different groups of contributors – authors, publishers and readers. Incorporating the full publishing lifecycle into a single system, striving for integration and collaboration between the different participants, brings great benefits. In this paper, we take the benefits of the social web for well-established[2] and accepted and focus on the more novel semantic aspects of the publishing realm. To this extent, we develop our work in the context of the Planetary eMath3.0 system (see [Koh+11] for an introduction). Notably, the Planetary framework implements the architecture introduced in [Dav+10] for publishing its documents as XHTML+RDFa+MathML, enabling interactive semantic services.

In our work on units and quantities, we have concentrated on setting the necessary technological foundation, hence building on the languages introduced in section 2.1 to select and enhance the authoring and interaction aspects.

3 Semantic Units – Idea Outline

In order to understand how a semantic representation of units and quantities will integrate with the publishing flow of our framework of choice, one first needs to pinpoint what they comprise and how they are *represented*.

A computational *semantic entity* is an object with explicit *structure*, representable in a machine-understandable form, and denoting a corresponding real-world entity. The denotation is usually encoded via a machine-readable ontology. This definition is directly applicable to semantic units and quantities, which are exactly the machine-readable representations of their physical counterparts.

For the *representation* we choose OpenMath, since it encompasses units through modular ontologies, called Content Dictionaries (CDs) [Col09], which enable extensibility through the creation of new such ontologies that can add new symbols or simply through the extension of the existing unit ontologies/CDs.

[2] For mathematics, including the mathematical foundations of science and engineering, see, e.g., the PlanetMath free encyclopedia [Pla] and the Polymath wiki/blog-based collaboration effort [Bar10].

As a running example for this paper, we consider a semantic representation of the physical ***quantity*** $\boxed{100\,\mathrm{km/h}}$; one possibility to represent it in OpenMath is[3]:

```
<OMA>
    <OMS cd="arith1" name="times"/>
    <OMI>100</OMI>
    <OMA>
       <OMS cd="arith1" name="divide"/>
       <OMA>
         <OMS cd="units_ops1" name="prefix"/>
         <OMS cd="units_siprefix1" name="kilo"/>
         <OMS cd="units_metric1" name="metre"/>
       </OMA>
       <OMS cd="units_time1" name="hour"/>
    </OMA>
</OMA>
```

Listing 1.1. OpenMath representation of $\boxed{100\,\mathrm{km/h}}$

4 State of the Art

We review the relevant prior work involving units and quantities in the context of semantic publishing. Note that we do not cover the publishing dimension itself, as it is a stand-alone framework level, independent of the processed content.

4.1 Representation

The semantic publishing aspect of units in scientific documents has not yet accumulated a sizable body of prior work. Previous research has been mainly concerned with the standardization of unit and quantity representation which is far from complete (not covering every unit occurrence possibility) or sufficiently machine comprehensible. There is a number of units-related semantic web ontologies: The authors of the Measurement Units Ontology [BP09] review a number of ways of representing units in RDF. The SWEET ontology (Semantic Web Earth and Environmental Terminology [Swe; RP05]) is particularly remarkable for linking units to the fields of science where they occur. A general weakness of RDF/OWL unit ontologies is, however, that the computation of derived units (and thus unit conversion) cannot be described in a straightforward way (and is rarely done).

For OpenMath, a representation of units and quantities has been proposed (cf. [DN03]), and several CDs covering common units have been provided. The

[3] This is one out of several ways of representing units (cf. [DN03]). For a detailed description of the XML schema see section 3.1.2 of [Bus+04]

in-depth analysis of the prospective representations of units and their dimensions that [DN03] proposes (taking into account the pros and cons of each approach) allows for a broader view to the multitude of semantic publishing possibilities. The two most significant sets of OpenMath unit CDs have been developed by James Davenport and Jonathan Stratford [SD08] and Joseph Collins [Col09], respectively. The former are remarkable for their explicit representation of conversion rules (see also Section 4.3). The latter ones provide a standards-compliant implementation of SI[4] quantities and units, providing strong insight on the concepts of *quantity* and *unit* and on the prospects of capturing more of their semantics in the representation.

4.2 Authoring

In "pre-semantic" environments, such as LaTeX, there are first approximations of content-oriented macros that represent units. A prominent example is the LaTeX package *SIunits* [Hel] which covers the full range of base and derived units in the SI system, as well as SI prefixes, a range of widely accepted units external to SI and a couple of generic mechanisms for creating custom author-specified unit constructs. The package enables a large set of abbreviative commands, which are internally built up from the compositional application of atomic building blocks. In this sense, the authoring process via *SIunits* is *nearly semantic* on the interface level, but *entirely presentational* on the output side.

Still, all major semantic authoring systems (e.g. the semantic LaTeX extensions sTeX [Koh08], SALT [Gro+07], the Ontology Add-in for Microsoft Office Word [Fin+10], or the semantic content management system PAUX [PAU]) have so far neglected the specific use case of units. This can be partially explained by the lack of a widely agreed standard representation, as well as different primary development foci – mathematics for sTeX, rhetorical structures for SALT, life sciences terminology for the Word ontology add-in, and educational texts from areas unrelated to physics, such as law, for PAUX. Notably, sTeX could, in principle, support units already, as its wide coverage of the conceptual model of OpenMath and its generic mechanism for defining new symbols and concepts could easily be utilized for the specification of the relevant unit and quantity symbols. Section 5 presents how we have done that in a way that does not disrupt existing LaTeX authoring practices. While LaTeX is commonly used in mathematics, science, and engineering, our solution is unlikely to appeal to life scientists, where Microsoft Office Word is more widely used; however, we leave unit support for word processors to future work.

4.3 Interaction

Applications taking advantage of the semantic publishing of units and their quantities using OpenMath have also been experimented with by various authors, albeit the lack of authoring support. The unit conversion service [Str08; SD08]

[4] The International System of Units [Sib]

by Jonathan Stratford, which users can easily extend by uploading new Content Dictionaries (CDs) with new units and conversion rules, provides a good example of the power of semantically annotated units. Besides the implementation of such a service, Stratford's research also identifies the difficulties of unit conversion and the limitations of OpenMath's current state with regard to unit representation.

Stratford's conversion service is interactive in that users can enter quantities into a web form and upload definitions of new units. We have additionally made it interactively accessible from web documents that contain MathML formulas with OpenMath annotations, as created by the publishing pipeline explained in section 2.2 (cf. [GLR09]). This interaction with units in publications has, however, remained a proof of concept so far, as *producing* suitably annotated documents required manual authoring of quantity expressions in OpenMath XML markup – a barrier that we are trying to overcome with the work presented in this paper.

5 Semantic Authoring of Units and Quantities

We have revised the available methods and technologies and established that a semantic authoring support for units does not formally exist at present. Consequently, we set out to make the first steps towards extending one of the more prepared software solutions, namely sTeX, with a special authoring module for units, by building on the existing pre-semantic toolbox of the *SIunits* LaTeX package. sTeX [Koh08] is essentially a collection of LaTeX packages that offer semantic macros. sTeX can be translated into XML markup using LaTeXML [Mil] bindings, thus enabling easier subsequent processing – including semantic web publishing (cf. [Dav+10]). Our units extension follows a similar approach[5].

As described in section 4.2, *SIunits* provides an sTeX-like content authoring interface. For our running example, we are interested in authoring $\boxed{100\,\mathrm{km/h}}$ in order to create the content representation shown in Listing 1.1. There are many ways to author the representation in LaTeX, e.g. via `$\textrm{100\,km/h}$`. The *SIunits* package makes the process less ad-hoc by focusing on the content and factoring out the presentational quirks, in the form of package options. Hence, one would instead write the more semantic `\unit{100}{\kilo\metre\per\hour}`. It is interesting to observe that a completely different motivation than ours, namely to provide a convenient and centralized interface to control the *presentation* of the unit entities on a document level, leads to the essentially same result which we desire – a *semantics-oriented* authoring interface.

In our effort to leverage this functionality, we first created a LaTeXML binding for the *SIunits* package. It helped us to pinpoint the semantic map between the interface and the OpenMath representation and provided a non-invasive semantic enrichment for LaTeX documents based on the package. Next, we use the gained understanding in building a native sTeX module for units, roughly

[5] The SIunits bindings and sTeX extension will be released in the respective bundles (the arXMLiv binding library and the sTeX package on CTAN) with the authors' strong committment to free software licenses compatible with the originals.

based on the *SIunits* interface. Table 1 shows a small snippet comparing the different stages. One easily notices the abbreviative power of the sTEX approach, which hides the verbose and overly complex binding declaration under its hood, exposing the author to a controlled LATEX vocabulary and facilitating reuse.

Language	Definition	Semantics
LATEX	`\newcommand{\kilo}{\ensuremath{\mathrm{k}}}` `\newcommand{\metre}{\ensuremath{\mathrm{m}}}`	✗

LATEXML	`DefConstructor('\kilo{}','`
	`<ltx:XMApp>`
	`  <ltx:XMTok meaning="prefix" cd="units_ops1"/>`
	`  <ltx:XMTok meaning="kilo" cd="units_siprefix1">`
	`  k`
	`  </ltx:XMTok>`
	`  #1`
	`</ltx:XMApp>');`
	`DefConstructor('\metre','`
	`<ltx:XMTok meaning="metre" cd="units_metric1">`
	`  m`
	`</ltx:XMTok>');`
		✓

sTEX	`\symdef[name=kilo,cd=units_siprefix1]{kiloPX}{\mathrm{k}}`
	`\symdef[name=metre,cd=units_metric1]{metre}{\mathrm{m}}`
	`\symdef[name=prefix,cd=units_siprefix1]{prefixFN}{}`
	`\symdef{kilo}[1]{\mixfixii{}{\kiloPX}{\prefixFN}{#1}{}}`
		✓

Table 1. Definitions for \kilo\metre, typeset as 'km'

6 Interaction with Units and Quantities

Given the provisions for authoring support, we move to the added-value benefits one could reap from interacting with a published document. This section details relevant use cases and explains the prerequisites that are already available.

6.1 Unit (System) Preference Service

A concrete scenario for a prospective service that would take advantage of semantically published papers, based on the ideas from section 3, can be evolved on top of common published material like *cooking recipes*. These provide a good use case thanks to the high density of units and quantities they contain. Moreover, the physical quantities are restricted to a small subset (quantity/mass related units) including special types of *units* [21c] which are not formally defined and might prove to be misleading:

$$1\ \textbf{teaspoon (tsp)} \approx 5\ \textit{millilitres (mL)}$$
$$1\ \textbf{cup} \approx 250\ \textit{millilitres (mL)}$$

The idea of the *unit (system) preference* service is to allow the user/reader to choose a preferred system of units (e.g. imperial, metric) or simply preferred types of units (e.g. "minutes" instead of "hours", "kilogrammes" instead of "grammes") for the representation of physical quantities and then seamlessly adapt the document to these preferences. This can only be achieved at the end of the semantic publishing pipeline, since the process requires the technologies described in sections 2 and 4 for the *representation* and *authoring* parts. Once these prerequisites have been met, one can embed interactive scripts into the published document (here: XHTML with OpenMath-annotated MathML formulae), which invoke a web service for any computation. In our implementation, the JOBAD (Javascript API for OMDoc-based Active Documents) framework [GLR09] provides for client-server communication and manipulation of the document. Figure 1 visualizes the workflow.

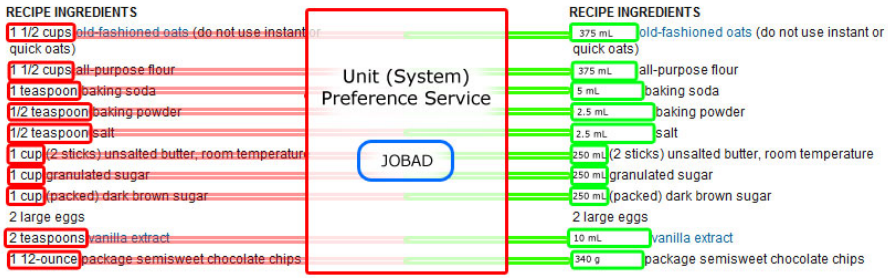

Fig. 1. Workflow for Chocolate Chip Cookies recipe [Crc]

6.2 Prospective Services based on Semantically Published Units

Having described in detail one service that enhances the user experience by publishing units semantically, we now list further potential services and applications that the same technology could enable:

- **Mapping Natural Sciences Concepts to their respective Units:** defining Content Dictionaries that would describe the connection of units to general natural sciences concepts like *force* (measured in Newtons: $N = \frac{kgm}{s^2}$ or any variant of the ratio) or *energy* (measured in Joules: $J = Nm = \frac{kgm^2}{s^2} = ...$) and plenty of other examples. The interconnection of concepts in sciences: $Energy = Force \times displacement$ can further enable scientific formula "spell checking" which might prove to be of great value to physicists, astronomers and many others.
- **Unknown Unit Lookup:** In theoretical scientific papers authors usually use abbreviations for concepts (e.g. N for *Newtons* – the unit for *force*) without mentioning anything about units/dimensions, which might turn out

to be difficult for the readers who would be interested to know, for example, the order of measurement (magnitude) for the unknown physical quantities and also a (small) description of the respective concept (e.g. Pa is the unit for *pressure*). Defining a generic way in which semantics can be added to such unknown symbols will enable showing/hiding units for expressions/formulas.
- **Unit and Quantity Semantic Search:** a library-level service that would allow searching for units by their type, name and magnitude and return the relevant results independently of the measuring standard of the occurrences in the paper (e.g. imperial or metric) and also independent of their form (N or $\frac{kgm}{s^2}$).[6]
- **Quantity and Unit's Magnitude Manipulation:** a document interaction service that is able to transform for example $100N \rightarrow 0.1kN$ or $0.1 \times 10^3 N$ or $0.1 \times 10^3 \frac{kgm^2}{s^2}$. This can be useful when it comes to simplifying representations and adapting them consistently to a certain type of magnitude (for example *all occurrences of force expressions should have their unit represented in kN*).

As detailed at the beginning of this paper, having a standard, uniform understanding of units and quantities can prevent hazards and even eliminate entire compatibility check processes in industry. The presented list of prospective enabling technologies shows only a few of the numerous opportunities of interacting with units and quantities in semantically published documents and serves as strong motivation for future research in this direction.

7 Enabling Semantic Units in Legacy Corpora

The arXMLiv corpus is the ideal environment for the identification of units and quantities since it contains a collection of more than 600,000 scientific publications. It is based on Cornell University's arXiv e-Print archive [Arx] originally typeset in LaTeX, converted to XML in order to achieve easy machine-readability, partial semantics recovery and clear separation of document modalities such as natural language and mathematical expressions [Sta+10]. Currently, the project has achieved a successful conversion rate of nearly 70% to a semantically enriched XHTML+MathML representation, natively understandable by modern web browsers [Koh+08].

A proof-of-concept check, performed via the arXMLiv build system (see [Sta+10]) revealed roughly 150 arXiv articles using the *SIunits* package, with an outlook for close to tripling the number when considering sibling packages such as *units* and *SIunitx*. This gives our work on creating a semantic binding for *SIunits* an even stronger benefit, as we can directly and non-invasively enrich legacy publications, putting them one step further on the path to semantic publishing. An additional, mid-term benefit is the opportunity to build a linguistic *Gold Standard* for units; we created both legacy (to presentational MathML) and semantic (to OpenMath) bindings in order to provide a raw, presentational

[6] In contrast, state-of-the-art scientific publication search services, such as Springer's LaTeX search [Spr], do not support the semantics of units.

output and its annotated, semantic counterpart. Having both as a basis, unit spotters can then be developed using methods of Computational Linguistics and Machine Learning, further enriching the arXMLiv corpus.

Such enhancements not only enable the interactive services of semantic publishing on legacy corpora, but also provide a tempting outlook to the development of an ecosystem of linguistic analysis modules, which can draw on the captured semantics of units and quantities, as originally envisioned by the LaMaPUn project [Gin+09].

8 Conclusions and Future Work

Units and quantities are sufficiently wide-spread and important to not be disregarded from the context of semantic documents. Unfortunately, by now, there have been only isolated approaches (see section 4) to exploit the semantic power of units. Also considering the wide range of existing unit types and representations, makes it almost impossible to identify and semantically enrich all of them, especially when we are talking about occurrence contexts as unrelated as cooking recipes, medical prescriptions, technical documents or scientific papers.

Through the separation of the semantic publishing process for units we emphasized the importance of three major components: *representation*, *authoring* and *interaction*, detailing technologies that can improve each of them. Moreover, by providing a cooking recipe interaction use case and also a series of further potential services and applications on top of semantically published units, we contribute means of better manipulation and interpretation of *units and quantities* to the Semantic Publishing Industry and to legacy corpora.

Acknowledgments: The authors would like to thank Michael Kohlhase for his extensive support and advice regarding the writing of this paper, the anonymous peer reviewers for their extensive helpful suggestions, and Anton Antonov for writing the LaTeXML bindings for the *SIunits* LaTeX package.

References

[21c] *Code of Federal Regulations – Food and Drugs*. 2004. URL: http://edocket.access.gpo.gov/cfr_2004/aprqtr/21cfr101.9.htm.

[Air] *Aviation Safety – Air Canada Accident Report*. URL: http://aviation-safety.net/database/record.php?id=19830723-0 (visited on 10/25/2010).

[Arx] *arxiv.org e-Print archive*. URL: http://www.arxiv.org.

[Aus+10] R. Ausbrooks et al. *Mathematical Markup Language (MathML) Version 3.0*. Recommendation. W3C, 2010. URL: http://www.w3.org/TR/MathML3.

[Bar10] M. J. Barany. "'[B]ut this is blog maths and we're free to make up conventions as we go along': Polymath1 and the modalities of 'massively collaborative mathematics'". In: *WikiSym*. Ed. by P. Ayers and F. Ortega. ACM Press. 2010.

[BP09] D. Berrueta and L. Polo. *Measurement Units Ontology*. Nov. 9, 2009. URL: http://forge.morfeo-project.org/wiki_en/index.php?title=Measurement_Units_Ontology&oldid=12301.

[Bus+04] S. Buswell et al. *Open Math Standard 2.0*. Tech. rep. OpenMath Society, 2004. URL: http://www.openmath.org/standard/om20.

[Col09] J. B. Collins. "OpenMath Content Dictionaries for SI Quantities and Units." In: *Calculemus/MKM*. Ed. by J. Carette et al. Vol. 5625. LNCS. Springer, 2009, pp. 247–262.

[Crc] *Cooking.com – "Giant Chocolate Chip Cookies"*. URL: http://www.cooking.com/recipes-and-more/recipes/Giant-Chocolate-Chip-Cookies-recipe-5112.aspx (visited on 03/05/2011).

[Dav+10] C. David et al. "Publishing Math Lecture Notes as Linked Data". In: *The Semantic Web: Research and Applications (Part II)*. Ed. by L. Aroyo et al. LNCS 6089. Springer, 2010, pp. 370–375. arXiv:1004.3390v1 [cs.DL].

[DN03] J. H. Davenport and W. A. Naylor. *Units and Dimensions in OpenMath*. 2003. URL: http://openmath.org/documents/Units.pdf.

[Fin+10] J. L. Fink et al. "Word add-in for ontology recognition: semantic enrichment of scientific literature". In: *BMC Bioinformatics* 11.103 (2010).

[Gin+09] D. Ginev et al. "An Architecture for Linguistic and Semantic Analysis on the arXMLiv Corpus". In: *Applications of Semantic Technologies (AST) Workshop at Informatik*. 2009. URL: http://www.kwarc.info/projects/lamapun/pubs/AST09_LaMaPUn+appendix.pdf.

[GLR09] J. Giceva, C. Lange, and F. Rabe. "Integrating Web Services into Active Mathematical Documents". In: *MKM/Calculemus Proceedings*. Ed. by J. Carette et al. LNAI 5625. Springer, 2009, pp. 279–293.

[Gro+07] T. Groza et al. "SALT – Semantically Annotated LaTeX for Scientific Publications". In: *The Semantic Web: Research and Applications*. Ed. by E. Franconi, M. Kifer, and W. May. LNCS 4519. Springer, 2007, pp. 518–532.

[Hel] M. Heldoorn. *The SIunits package: Consistent application of SI units*. URL: http://mirror.ctan.org/macros/latex/contrib/SIunits/SIunits.pdf (visited on 03/13/2011).

[Koh08] M. Kohlhase. "Using LaTeX as a Semantic Markup Format". In: *Mathematics in Computer Science* 2.2 (2008), pp. 279–304. URL: https://svn.kwarc.info/repos/stex/doc/mcs08/stex.pdf.

[Koh+08] M. Kohlhase et al. "MathWebSearch 0.4, A Semantic Search Engine for Mathematics". manuscript. 2008. URL: http://mathweb.org/projects/mws/pubs/mkm08.pdf.

[Koh+11] M. Kohlhase et al. "The Planetary System: Web 3.0 & Active Documents for STEM". In: accepted for publication at ICCS 2011 (Finalist at the Executable Papers Challenge). 2011. URL: https://svn.mathweb.org/repos/planetary/doc/epc11/paper.pdf.

[KR09] M. Kohlhase and F. Rabe. "Semantics of OpenMath and MathML3". In: *22nd OpenMath Workshop*. Ed. by J. H. Davenport. 2009. URL: http://kwarc.info/kohlhase/papers/om09-semantics.pdf.

[Lan11] C. Lange. "Ontologies and Languages for Representing Mathematical Knowledge on the Semantic Web". accepted by Semantic Web Journal. 2011. URL: http://www.semantic-web-journal.net/content/new-submission-ontologies-and-languages-representing-mathematical-knowledge-semantic-web.

[Mar] *CNN – "NASAs metric confusion caused Mars orbiter loss"*. 1999. URL: http://articles.cnn.com/1999-09-30/tech/9909_30_mars.metric_1_mars-orbiter-climate-orbiter-spacecraft-team?_s=PM:TECH (visited on 10/29/2010).

[Mil] B. Miller. *LaTeXML: A LaTeX to XML Converter*. URL: http://dlmf.nist.gov/LaTeXML/ (visited on 03/03/2011).

[Oxf] *Oxford English Dictionary. "unit" definition*. URL: http://dictionary.oed.com/entrance.dtl (visited on 10/29/2010).

[PAU] *PAUX Technologies*. URL: http://paux.de (visited on 10/10/2010).

[Pla] *PlanetMath*. URL: http://planetmath.org (visited on 01/06/2011).

[RP05] R. G. Raskin and M. J. Pan. "Knowledge representation in the semantic web for Earth environmental terminology (SWEET)". In: *Computers & Geosciences* 31 (2005), pp. 1119–1125.

[SD08] J. Stratford and J. H. Davenport. "Unit Knowledge Management". In: *Intelligent Computer Mathematics*. Ed. by S. Autexier et al. LNAI 5144. Springer, 2008, pp. 382–397.

[Sib] *The International System of Units (SI) 8. Edition*. Bureau International des Poids et Mesures, 2006. URL: http://www.bipm.org/utils/common/pdf/si_brochure_8_en.pdf.

[Spr] Springer, ed. *LaTeX Search*. URL: http://www.latexsearch.com (visited on 04/16/2011).

[Sta+10] H. Stamerjohanns et al. "Transforming large collections of scientific publications to XML". In: *Mathematics in Computer Science* 3.3 (2010): *Special Issue on Authoring, Digitalization and Management of Mathematical Knowledge*. Ed. by S. Autexier, P. Sojka, and M. Suzuki, pp. 299–307. URL: http://kwarc.info/kohlhase/papers/mcs10.pdf.

[Str08] J. Stratford. *Creating an extensible Unit Converter using OpenMath as the Representation of the Semantics of the Units*. Tech. rep. 2008-02. University of Bath, 2008. URL: http://www.cs.bath.ac.uk/pubdb/download.php?resID=290.

[Swe] *Semantic Web for Earth and Environmental Terminology (SWEET)*. NASA. URL: http://sweet.jpl.nasa.gov/ (visited on 08/22/2010).

[Usm] *US Metric Association "Unit Mixups" article*. URL: http://lamar.colostate.edu/~hillger/unit-mixups.html (visited on 10/25/2010).

Exploring the Generation and Integration of Publishable Scientific Facts Using the Concept of Nano-publications

Amanda Clare[1,3], Samuel Croset[2,3] (croset@ebi.ac.uk), Christoph Grabmueller[2,3], Senay Kafkas[2,3], Maria Liakata[1,2,3], Anika Oellrich[2,3], and Dietrich Rebholz-Schuhmann[2,3]

[1] University of Aberystwyth
[2] European Bioinformatics Institute
[3] All authors contributed equally to this work

Abstract. Publication formats are being sought that facilitate automatic processing and knowledge integration and are better suited to the current pace of research. Here we present an infrastructure for producing and consuming minimal publishable units, nano-publications, directly from a researcher's electronic notes or manuscripts which allow the integration of multiple resources. We describe a feedback loop resulting from the use of nano-publications, give a detailed example, and explain how this can be combined with existing web technologies.

1 Introduction

With the ever growing amount of scientific literature the automatic analysis of scientific content has become crucial. As part of the research life-cycle, researchers constantly need to retrieve relevant documents, pull out facts, reuse and reference them. Yet, currently many scientific facts are 'buried' in the plethora of information contained in traditional scientific publications. Repositories, such as PubMed[4], store electronic versions of scientific publications but the scientific facts they contain are still not available for automated processing. Research in recent years has looked at the enhancement of scientific publications with semantic meta-data in order to facilitate information retrieval and information extraction from them. Numerous initiatives from the publishing world have been launched to address this task, such as Structured Digital Abstracts [3], The Royal Society of Chemistry (RSC) Project Prospect[5] and UKPMC[6], a major initiative focusing on the integration and alignment of literature with current knowledge resources and databases. An example system for the enhancement of documents with content from a variety of external resources is Utopia Documents [1]. Meanwhile, scientists have begun to seek more rapid and interactive ways of airing their findings and retrieving the findings of others, through media such

[4] http://www.ncbi.nlm.nih.gov/pubmed/
[5] http://www.rsc.org/Publishing/Journals/ProjectProspect/
[6] http://ukpmc.ac.uk/

as blogs and wikis. Wikis are collaborative tools, enabling users to collect, share and edit information while blogs are incremental content management systems enabling rapid publication of information. In traditional wiki systems and blogs, accessing, querying, retrieving and aggregating data is difficult since the knowledge is represented in unstructured form. These issues induced the emergence of semantic wiki systems like Semantic MediaWiki and DBpedia and commercial or semi-commercial web content providers (e.g. Apture). However, such tools are currently not aimed at scientists and do not offer the precision and level of detail that scientists need to make their work unambiguous and available in a machine-readable, reusable form to others. Importantly, it is not easy to receive credit for statements on wikis or blogs, or to cite the information therein as one would do with a standard publication. The nano-publication (NP) [6] has been proposed as a new form of academic publishing. Unlike other initiatives for linking shared statements in the literature [7], a NP is defined as a citable unit containing a set of annotated statements which capture knowledge in the form of Resource Description Framework (RDF) triples, representing three concepts (subject, predicate, object) [4]. The RDF graph emerging from the triples can be identified with a name, a procedure which was coined as a 'named graph' [2]. Despite the establishment of concepts such as NPs and named graphs, no unique way has been identified to facilitate the integration of all possible ways of publishing in a fast and reliable way. Here, we demonstrate how our interpretation of NPs, named graphs, knowledge resources and existing web tools can be combined to facilitate the integration of the diverse types of publishing and potentially lead to the discovery of new knowledge.

2 Practical Example of NPs: Feedback Looping

In this paper, we explore the generation and use of NPs [4] by means of a concrete example. We define a NP as a set of one or more RDF statements which assert some knowledge in the field of expertise of the person publishing, or demonstrate an endorsement of a statement by the latter. While the RDF annotations may be automatically obtained using text mining methods, they will have been manually approved and collected in a set, constituting a new object, by the author of the NP. In addition, it is important that a NP can be properly cited and its origin (provenance) traced. Therefore, a NP should include an identifier, e.g, digital object identifier (DOI), while provenance should include the author of the NP, and the origin of each of the statements included therein. Based on the above, we propose a model for NPs, which is summarised in Figure 1.

The model is cyclic and based on dynamic interaction between users and the machine. It consists of three steps, focusing on benefits for creators and users of NPs. Firstly, researchers create the NP and offer it to the broader community (see Figure 2). The second step involves machine consumption of the data generated by users. Machines can integrate data from multiple sources as long as they are represented in a common format with explicit semantics (RDF). For example, statements generated by a user can be integrated with statements coming from

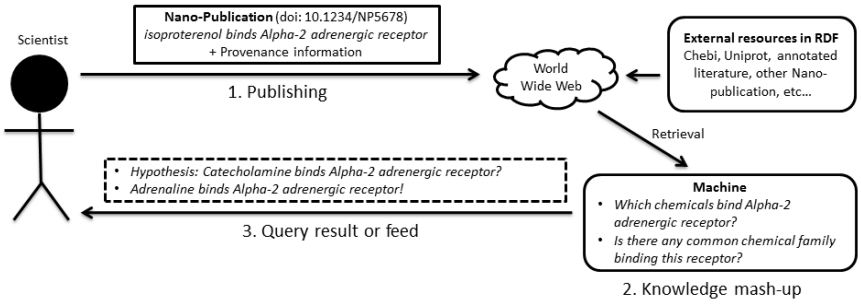

Fig. 1. A model for nano-publishing.

UniprotKB[7] or existing literature. Computers can also combine related NPs and generate new hypotheses. The third step of the model is the user's reward: feeds generated from the data integration done by the machine in the second step. The author of the NP will receive relevant information, based on previously published assertions processed by the machine. The hypotheses can be evaluated, rejected or validated by data added by the user, leading to a new NP, as described in step one. The dynamic human-computer interaction allows NP writers to access relevant information tailored to individualized retrieval, which they can enrich for the benefit of the community. As NPs are uniquely identified by a DOI, they can be cited and further used in any type of publication.

3 Working Example of NPs Using Semantic Wikis

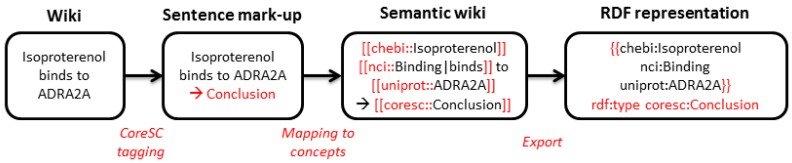

Fig. 2. General example of NP creation from a scientific statement contained in a wiki.

We propose that the route to a scientific NP can be facilitated by enabling the annotation of scientific notes or blogs at multiple levels of detail. The author of the notes can then package together aspects of their notes as a scientific publication. We have created a prototype of a tool for the open source wiki MediaWiki, allowing the user to manually annotate a scientific document with automated markup; multiple tags are allowed at the sentence level. This tool is aimed at scientists who use MediaWiki as an electronic lab notebook environment. Scientists could post a set of their annotated sentences as a NP. Additional annotation (manual or automated) of important entities and relations between entities can be provided for terms within these sentences. These can then be saved as triples linked to the NP. A mockup screenshot of entity-level markup can be found in Figure 3. The prototype of our tool can also model the scientific

[7] http://www.uniprot.org/help/uniprotkb

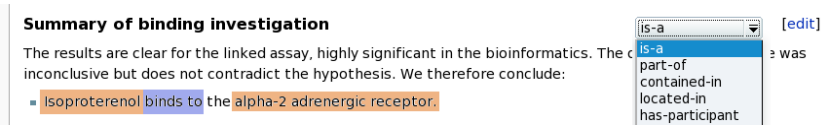

Fig. 3. Mock-up of annotation tool (automatic or manual) of significant entities and relations in a wiki used as a laboratory notebook.

discourse of the document or notes in terms of CoreSC [5], an annotation scheme successfully used to automatically recognise core scientific concepts in research articles. Thus we can retrieve the semantic context from which a NP has been generated (Result, Conclusion, Hypothesis, etc.). Once generated, the RDF form of the NP can then be exported and hosted on an external RDF hosting site. Figure 2 demonstrates this process.

The following example illustrates the format of NPs including a simple statement (Figure 4)[8]. The illustration is extended to show how a new hypothesis can be generated from the user's input.

```
@prefix chebi: <http://purl.org/obo/owl/CHEBI#> .
@prefix nci: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#> .
@prefix uniprot: <http://purl.uniprot.org/uniprot/> .
@prefix prv: <http://purl.org/net/provenance/ns#> .
@prefix coreSC: <http://www.sapientaproject.com/publications/coresc#>
{ chebi:CHEBI_6257 nci:C82888 uniprot:P08913 . }
  rdf:type prv:DataItem, coreSC:conclusion ;
  prv:createdBy [ rdf:type prv:DataCreation ;
    prv:performedAt "2009-07-10T12:00:00Z"^^xsd:dateTime ;
    prv:performedBy [a foaf:Person; foaf:name "Ben Huxley"] ;
    prv:accessedResource <http://www.mysite.com/mywiki/mydoc48#sentence65> ] ;
  _:uri doi:10.1234/NP5678 .
```

Fig. 4. RDF representation of exemplary conclusion.

A researcher derives the conclusion: "isoproterenol binds to the Alpha-2 adrenergic receptor" and decides to create a NP corresponding to the conclusion. In the semantically enriched statement, *isoproterenol* is mapped to the ChEBI ontology (ChEBI)[9], the receptor to UniprotKB, and the action *binds* to one of the terms of the NCI Thesaurus[10]. Automated URI mappings can be achieved with services like the NCBO BioPortal[11]. Once the data has been transformed to RDF, the NP is available for further processing and integration into the network of linked data.

```
select ?chemical where {
  ?chemical rdf:type nci:C48807 .
  ?chemical nci:C82888 uniprot:P08913 .
}
```

Fig. 5. Query statement.

The example in Figure 5 illustrates how a generic query can be generated from the NP presented above. The query aims to retrieve other chemicals binding the receptor and is run over a knowledge base. The result is a list of known chemicals binding P08913. The adrenaline molecule (chebi:CHEBI_33568) will appear in the list and could be reported to the user. Applying a reasoner will reveal that adrenaline and *isoproterenol* are both members of the catecholamine family (chebi:CHEBI_33567). From this observation and in absence of any other information present in the databases, the following

[8] For provenance terminology and concepts we follow: http://trdf.sourceforge.net/provenance/ns.html
[9] http://www.ebi.ac.uk/chebi/
[10] http://ncit.nci.nih.gov/
[11] http://bioportal.bioontology.org/

hypothesis can be sent to the creator of the NP: “Do compounds from the catecholamine family bind the P08913 receptor?”. The newly formed hypothesis can give rise to new experiments and be published in return (Figure 6), citing both the NP above and ChEBI, which are the basis of the hypothesis.

```
{ CHEBI:33567 nci:C82888 uniprot:P08913 } rdf:type prv:DataItem ;
  prv:createdBy [ rdf:type prv:DataCreation ;
    prv:performedAt "2009-07-10T12:00:00Z"^^xsd:dateTime ;
    prv:performedBy [a foaf:Person; foaf:name "Dan Brickley"] ;
    prv:usedData _:otherNanopub , _:chebiWeb ] ;
  _:uri doi:10.1234/OQ6789 .

_:otherNanopub rdf:type prv:DataItem ;
  prv:retrievedBy [ rdf:type prv:DataAccess ; prv:accessedResource doi:10.1234/NP5678 ] .
_:chebiWeb rdf:type prv:DataItem ;
  prv:retrievedBy [ rdf:type prv:DataAccess ;
  prv:accessedResource <http://www.ebi.ac.uk/chebi/displayAutoXrefs.do?chebiId=CHEBI:33568> ] .
```

Fig. 6. Potentially new NP.

4 Conclusion

We have described an infrastructure for creating NPs using web tools such as blogs and wikis, while integrating information from a number of external resources. We also demonstrated how the produced NPs can be extended via the integration of information from other knowledge resources through querying across available resources on a semantic layer. A researcher using this system can decide whether or not to confirm the result of a query resulting from their initial publication and publish it as a new NP. Enabling this strategy of publishing will not only facilitate the integration of diverse resources but also allow for fast and precise knowledge dissemination and retrieval.

References

1. Attwood, T.K., et al.: Utopia documents: linking scholarly literature with research data. Bioinformatics 26(18), i568–74 (2010)
2. Carroll, J.J., et al.: Named graphs, provenance and trust. International World Wide Web Conference (2005)
3. Gerstein, M., et al.: Structured digital abstract makes text mining easy. Nature 447(7141), 142 (2007)
4. Groth, P., Gibson, A., Stickler, P.: The anatomy of a nanopublication. Information Services and Use 30(1), 51–56 (2010)
5. Liakata, M., et al.: Corpora for the conceptualisation and zoning of scientific papers. In: International Conference on Language Resources and Evaluation (2010)
6. Mons, B., Velterop, J.: Nano-publication in the e-Science era. Workshop on Semantic Web Applications in Scientific Discourse (Jan 2009)
7. Passant, A., Ciccarese, P., Breslin, J., Clark, T.: SWAN/SIOC: aligning scientific discourse representation and social semantics. Workshop on Semantic Web Applications in Scientific Discourse (2009)

A Framework for Semantic Publishing of Modular Content Objects

Catalin David, Deyan Ginev, Michael Kohlhase, Bogdan Matican, Stefan Mirea

Computer Science, Jacobs University, Germany; http://kwarc.info/

Abstract. We present the Active Documents approach to semantic publishing (semantically annotated documents associated with a content commons that holds the background ontologies) and the Planetary system (as an active document player).
In this paper we explore the interaction of content object reuse and context sensitivity in the presentation process that transforms content modules to active documents. We propose a "separate compilation and dynamic linking" regime that makes semantic publishing of highly structured content representations into active documents tractable and show how this is realized in the Planetary system.

1 Introduction

Semantic publication can range from merely equipping published documents with RDFa annotations, expressing metadata or inter-paper links, to frameworks that support the provisioning of user-adapted documents from content representations and instrumenting them with interactions based on the semantic information embedded in the content forms. We want to propose an entry to the latter category in this paper. Our framework is based on *semantically annotated documents* together with semantic background ontologies (which we call the **content commons**). This information can then be used by user-visible, semantic services like program (fragment) execution, computation, visualization, navigation, information aggregation and information retrieval (see Figure 5). Finally a document player application can embed these services to make documents executable. We call this framework the **Active Documents Paradigm** (ADP), since documents can also actively adapt to user preferences and environment rather than only executing services upon user request. In this paper we present the ADP with a focus on the Planetary system as the document player (see Figure 1)

The Planetary system (see [Koh+11; Dav+10; Plab] for an introduction) is a Web 3.0 system[1] for semantically annotated document collections in Science, Technology, Engineering and Mathematics (STEM). In our approach, *documents published in the Planetary system become flexible, adaptive interfaces to a content commons* of domain objects, context, and their relations. The system achieves

[1] We adopt the nomenclature where Web 3.0 stands for extension of the Social Web with Semantic Web/Linked Open Data technologies.

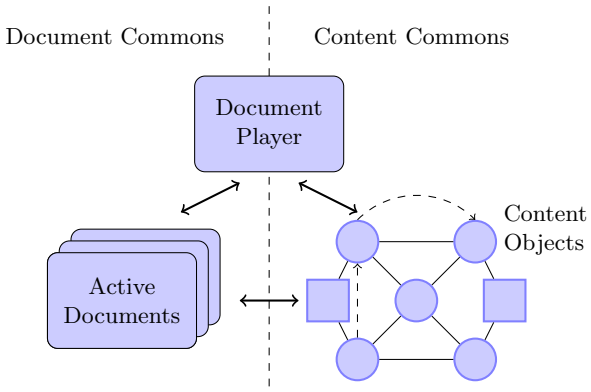

Fig. 1. The Active Documents Paradigm

this by providing embedded user assistance through an extended set of user interactions with documents based on an extensible set of client- and server side services that draw on explicit (and thus machine-understandable) representations in the content commons.

However, the flexibility and power designed into the active documents paradigm comes at a (distribution) cost: Every page that is shown to the user has to be assembled for the user in a non-trivial compilation process (which we call the **presentation** process) that takes user preferences and context into account. On the other hand, if the content is organized modularly, it can be reused across contexts. This presents a completely new set of trade-offs for publishing. One of them is that an investment in modular and semantic representational markup enhances reusability and thus may even lower the overall cost of authoring. We will explore another such trade-off in this paper: optimizing the distribution costs for modular content by "separate compilation".

In the next section we will look at the organization of the content presented to the user. This will constitute the conceptual backdrop against which we can discuss the issues involved in separate compilation and how we have solved them in the Planetary system.

2 Organization of Content/Narrative Structure

The Planetary system is intended as a *semantic publishing framework*, i.e. as a system providing the baseline capabilities needed for multiple specialized instantiations. We have shown the initial feasibility of the concept in a variety of publicly available case studies[2] ranging from pre-semantic archives of scientific literature [Arx], over a community-driven mathematical encyclopedia [Plac]

[2] Note that all of these are research systems under constant development, so your mileage may vary.

and the course system PantaRhei [Koh+], to a community portal of formal logics [Plaa]. As a consequence of this, we employ the general, modular knowledge structure depicted in Figure 2.

2.1 Levels of Content/Documents

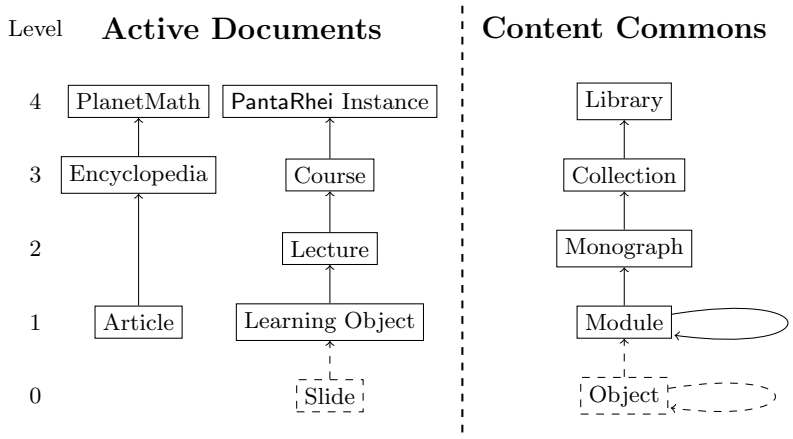

Fig. 2. Levels of Organisation of Content

The lowest level consists of atomic "modules"[3], i.e. content objects that correspond to small (active) documents dedicated to a single topic. For a course management system these might be learning objects (either as single modules or module trees), for an encyclopedia these would be the individual articles introducing a topic. Note that technically, we allow modules to contain (denoted by the arrows) other modules, so that larger discourse structures could be formed. For example, sections can be realized as modules referencing other modules of subsections, etc. The next level up is the level of "monographs", written works on a single subject that have a complete, self-contained narrative structure, usually by a single author or group of authors who feel responsible for the whole monograph. As a content object, a monograph is usually built up from modules, e.g. as a "module tree" that corresponds to sectioning structure of traditional books, but often also includes front and backmatter such as a preface, acknowledgements (both special kinds of modules), table of contents, lists of tables and figures, an index and references (generated from content annotations). Figure 3 shows course notes in the PantaRhei system, while other documents at the mono-

[3] The level of objects below modules consists of individual statements (e.g. definitions, model assumptions, theorems, and proofs), semantic phrase-level markup, and formulae. Even though it carries much of the semantic relations, it does not play a great role for the document-level phenomena we want to discuss here in this paper.

graph level are articles in a journal, or books in a certain topical section of a library.

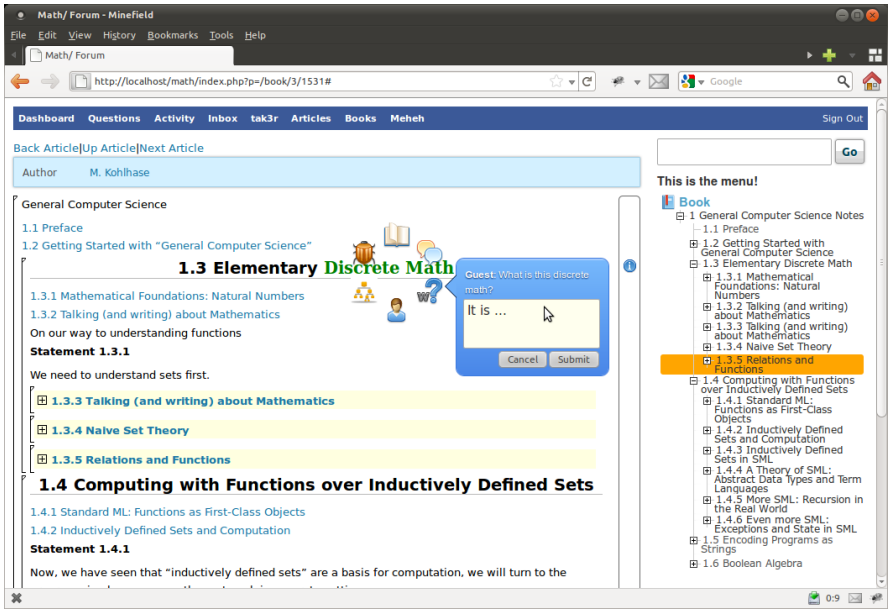

Fig. 3. A Monograph (Course Notes) in the Planetary system

Multiple monographs can be combined into collections, adding special modules for editorial comments, etc. Concrete collections in the document realm are encyclopedias, academic journals, conference proceedings, or courses in a course management system. Finally, the library level collects and grants access to collections, concrete, modern-day examples are digital libraries, installed course management systems, etc. In practice, a library provides a base URI that establishes the web existence of the particular installation. In the Semantic Web world, the library is the authority that makes its resources addressable by URLs.

2.2 Content Objects and their Presentations in Active Documents

To understand the differences between content objects and the documents generated from them in the presentation process, let us consider the example in Figure 4. Even though internally the content objects in Planetary are represented in OMDoc [Koh06], we will use the surface language sTEX[4] for the example, since this is what the author will write and maintain. sTEX is a variant of LaTeX

[4] We speak of an OMDoc surface language for any language that is optimized for human authoring, but that can be converted to OMDoc automatically.

that allows to add semantic annotations in the source. It can be transformed into OMDoc via the LaTeXML daemon [GSK11] for management in Planetary; see [Koh08] for details. We are using an example from a mathematical document[5] since content/presentation matters are most conspicuous there. In our experience, sTeX achieves a good balance (at least for authors experienced with LaTeX) conciseness and readability for mathematical documents. In particular, since sTeX documents such as the one in Figure 4 can be transformed to PDF via the classical `pdflatex` for prototyping and proofreading. The semantic editing process can further be simplified by *semantic document development environments* like sTeXIDE [JK10], which provides edit-support services lilke semantic syntax highlighting, command completion/retrieval, and module graph management.

```
\begin{module}[id=binary-trees]
  \importmodule[\KWARCslides{graphs-trees/en/trees}]{trees}
  \importmodule[\KWARCslides{graphs-trees/en/graph-depth}]{graph-depth}
  ...
  \begin{definition}[id=binary-tree.def,title=Binary Tree]
    A \definiendum[binary-tree]{binary tree} is a \termref[cd=trees,name=tree]{tree}
    where all \termref[cd=graphs-intro,name=node]{nodes}
    have \termref[cd=graphs-intro,name=out-degree]{out-degree} 2 or 0.
  \end{definition}
  ...
  \begin{definition}[id=bbt.def]
    A \termref[name=binary-tree]{binary tree} $G$ is called \definiendumalt[bbt]{balanced}
    iff the \termref[cd=graph-depth,name=vertex-depth]{depth} of all
    \termref[cd=trees,name=leaf]{leaves} differs by at most by 1, and
    \definiendum[fullbbt]{fully balanced}, iff the
    \termref[cd=graph-depth,name=vertex-depth]{depth} difference is 0.
  \end{definition}
  ...
\end{module}
```

Definition 3.1.7: *(Binary Tree)*
A **binary tree** is a tree where all nodes have out-degree 2 or 0.

Definition 3.1.8: A binary tree G is called **balanced** iff the depth of all leaves differs by at most by 1, and **fully balanced**, iff the depth difference is 0.

Fig. 4. Content and Presentation of an Object in sTeX

The upper half of Figure 4 shows the content representation of a module on binary trees, and its presentation in Planetary is in the lower box. The first aspect that meets the eye is that the presentation process[6] adds the textual marker "**Definition 3.1.7**" which is not present in the content representation \begin{**definition**}[id=binary-tree.def,title=Binary Tree]. Note that there are (at least) four issues at hand here pertaining to the presentation of the text marker:

[5] Actually from a second-semester course on Computer Science [Koh] hosted in PantaRhei— an instance of the Planetary system that is optimized for active course notes and discussions.

[6] We disregard the presentation of formulae in content representation like OpenMath or content MathML into presentation MathML in this paper and refer the reader to [KMR08] for details.

1. The marker "**Definition**" is context-sensitive: The presentation of a Spanish text would have generated "**Definición**".
2. The number "**3.1.7**" is content-sensitive in a totally different way: it is determined by the document structure, here it is a consequence of being the seventh definition in the first section in chapter 3.
3. The "house style" of a journal might use a different font family for the whole textual marker, for the text of the definition, or add an end marker for a distinctive layout. For instance in mathematical publications, theorems are usually set in italics and proofs use a box on the right of the last line as an end marker.
4. Finally, the whole text marker may be left out altogether in some situations, where a less formal presentation is called for.

Note that all these considerations have to be taken into account when referencing objects like these definitions. More so, these dimensions combine into a unique multi-dimensional point, which identifies the exact presentation of a document fragment. A content reference **sref**{binary−tree.def} might be presented as "**Def. 3.1.7**", in the same context as above (again subject to language, house style, etc). Note that here the style (e.g. the keyword) and generated contextual locators (e.g. the number) of the referenced object determines the actual label of the reference[7]. We follow the context dimensions specified in [KK08, Chapter 3], but note that many of the phenomena involve a separate, publishing context dimension (e.g. "house style").

Another phenomenon related to referencing is induced by the term reference **termref**[cd=graphs,name=vertex]{node}, which identifies the phrase "node" as a technical term and links it to its defining occurrence by the symbol name (here vertex) and the module name (also called content dictionary; here graphs). The specified module must be accessible in the current module via the **importmodule** relation and must contain a definition that contains a definiendum with symbol name vertex. The content module in Figure 4 specifies a module/content dictionary with name balanced−binary−trees, whose first definition supplies a definiendum with name balanced−tree via the **twindef** macro, which is referenced in the second definition. Note that in the presentation process where term references are displayed e.g as hyperlinks to the definition the name-based semantic links have to be converted into regular URI references. For this presentational conversion to hyperlinks one utilizes not only the module tree structure (i.e. visibility relationship) but also the library context that provides the base of the URI.

Finally, note that some content objects contribute to the context of other objects higher up in the content hierarchy in Figure 2. A good example for this are the definienda discussed above. In SᴛEX, they trigger index entries that populate the backmatter of monographs that include the respective module. Section titles populate the frontmatter in a similar way. Concretely, we have a top-level index stub in the backmatter, which "builds" itself from the context. In a sense, the index is an abstract concept with volatile presentation, generated

[7] a rather peculiar notion of context when viewed from a content-only perspective

from the module tree with the help of the content commons, which answers what objects should be indexed.

3 Separate Compilation

We have seen above that the various contexts (conceptual/document/language) have a significant effect on the presentation. But observe that if all the context-dependent parts of the presentation can be generated (albeit laboriously), the content representations are context-independent and can be reused in different contexts. This makes the content representations very portable. Consider for instance the definitions in our example above. They have been reused not only in eight instances of the "General Computer Science II" course [Koh] in the years 2004-2011 (each time with different numbers due to additions or deletions of preceding material), but also in different courses, e.g. as a recap in a more advanced CS course (without definition marker). But these are not the only contexts: the Planetary system can generate "guided tours" (self-contained explanations adapted to the user's prerequisite knowledge) for any concept in a document. Clearly, we cannot reasonably pre-compute all necessary presentation variants.

Computationally, the described situation is analogous to (and in fact conceptually influenced by) the situation in software design, where large programs are broken up into reusable source modules. As source modules are re-used in many programs, it is important that compilers support a regime of "separate compilation and linking" to make software development tractable: if one of many software modules used in a program changes, only that one module has to be re-compiled and the whole program re-linked. The first factor that enables this is the observation that for compilation of a module only the (relatively stable) signatures[8] of modules it depends on are needed, not the (relatively change-prone) module implementations. The second factor is that source modules can be compiled into a form, where references to functions imported from other modules are left symbolic and can later be replaced by concrete static references by the linker. We will call such forms of modules **contextable**, since they are contextualized by the linker in the way described.

In the Planetary architecture semantic publishing consists of the transformation of content structures encoded in STEX to active documents encoded in XHTML+MathML+RDFa (see Section 3.2 for details). To foster reuse, and make the process tractable, we want to assemble active documents from reusable content modules much in the same way as assembling an executable program from source modules. To make the separate compilation analogy fertile for semantic publishing it is useful to look at the role of context in the separate compilation regime: source modules are compiled into a context-independent form, which is then contextualized by linking compiled modules together into a consistent configuration for a concrete program. In the next two sections we

[8] Signatures contain the names of functions/procedures, possibly their types, but not their implementations.

examine how the two factors identified as crucial for the separate compilation regime can be obtained in the context of semantic publishing.

3.1 Contextable Presentations

Just as in programming, *separate* compilation of content modules into active documents is impossible without contextable structures in the presentation. It is an original contribution of our work to introduce them in the document setting. Concretely, we make use of the XML styling architecture and computes context-independent presentations that can be contextualized later. For instance, the XHTML header for the first definition in Figure 4 has the following form.

```
<div id="binary−tree.def" class="omdoc−definition">
    <span class="omdoc−statement−header">
       <span class="omdoc−definition−number"/>7</span>
       <span class="omdoc−statement−title">Binary Tree</span>
    </span>
    ...
```

We can then add (house) style information via CSS:

```
span.omdoc−statement−header {font−weight:bold}
span.omdoc−statement−title:before {content:"("}
span.omdoc−statement−title:after {content:")"}
span.omdoc−definition−number:after {content: ":␣"}
span.omdoc−definition−number:before {content:"Definition␣"}
```

Note that the keywords are not represented explicitly in the XHTML presentation, but added by content declarations in the CSS. This allows to overwrite the default ones via cascaded language-specific CSS bindings, e.g. using

```
span.omdoc−definition−number:before {content:"Definición "}
```

Note furthermore, that the presentation process only adds preliminary statement numbers in the XHTML presentation (here the number 7, since the definition is the seventh statement in the module). In the Planetary system, these numbers are dynamically overwritten by values computed from the context; in our example "3.1.7". The case for references is similar; for the table of contents shown in Figure 3 the presentation generates

```
<div class="omdoc−expandableref">
    <span class="omdoc−ref−number">4</span>
    <span class="omdoc−reftitle">
       <a href="../computing−dmath.omdoc" class="expandable">
          Computing with Functions over Inductively Defined Sets
       </a>
  </span>
</div>
```

in the table of contents on the right and in the text. The CSS class omdoc−expandableref triggers the Planetary interaction that expands the references in place to get the expanding ToC and the main document that can be folded/unfolded via the Mathematica-style folding bars on the extreme left.

3.2 Supporting the Logistics of Separate Compilation: Dynamic Linking

The role of the module signatures (think C header files) is taken by sTEX module signatures, i.e. auxiliary files generated from sTEX content modules that excerpt the information about references, modules and their dependencies; see [Koh08] for details. This information is used to establish a mapping between the content commons and the document commons (see Figure 1) that can be queried for the semantic interaction services embedded into the active documents.

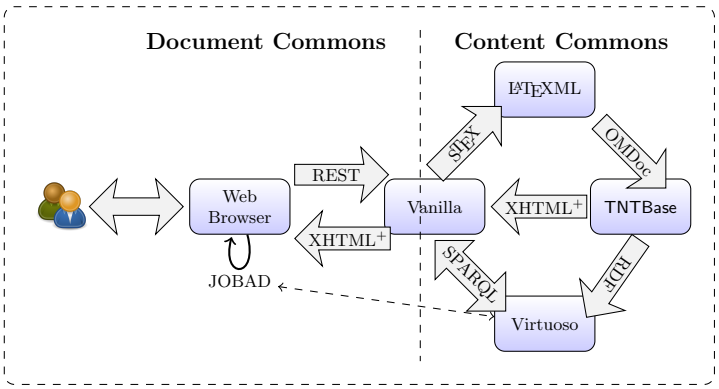

Fig. 5. The Planetary System Ecosystem

Actually, to understand the compilation and linking phases in Planetary, consider the system architecture in Figure 5. There, the document commons and content commons are layouted in a generic web browser and encapsulated versioned XML knowledge store TNTBase [ZK09] respectively. The Planetary system acts as an intermediary between these two:

1. sTEX is converted to OMDoc via the LaTeXML daemon [GSK11], this is then converted to XHTML+MathML+RDFa. Both transformations are highly dependent on notation information in the content commons, so they are under the control of the TNTBase system, which stores the content commons.
2. Planetary caches the contextable presentations that are generated by the TNTBase system. New presentations are requested from TNTBase whenever *a)* the content module in TNTBase has changed, and *b)* a user requests a to view the module.
3. Planetary hosts a triple store (Virtuoso) of structural metadata from the content commons that can be used for semantic services and document-level features, such as different views based on various selection criteria for an encyclopedia.
4. Finally, Planetary hosts structural information about the knowledge items at the different levels in Figure 2, used by the linker. In the examples in Fig-

ure 3 and Section 2.2, the numbering is linked into the contextable modules whenever a page is viewed, based on this information. Recall we need this *dynamic (i.e. view-time) linking* as modules are re-used in different document contexts.

4 Conclusion

In this paper we have explored the conceptual and practical decoupling and interaction of content and presentation in the active documents paradigm of semantic publishing. Our main focus rested on the interaction of content object reuse and context sensitivity of the presentation process. To make semantic publishing of highly structured content representations into active documents tractable we have developed a "separate compilation and dynamic linking" regime for transforming highly structured content representations into active documents. The concrete realization in the Planetary system hinges on the development of contextable pre-presentations that are contextualized at document load time.

While the basic architecture has been realized in the Planetary system, there is still a lot to explore in the active documents paradigm and its SCDL implementation. One crucial aspect is that while SCDL makes building active documents tractable, it also leads to the well-known "late binding problems" (aka "DLL Hell"), if modules change without adaptation of the dependent ones. We are currently working on an integration of an ontology-based management of change process [AM10] into the Planetary system (see [Aut+11]). This tries to alleviate late binding problems by analyzing the impacts of a change via the dependency relation induced by the semantic structure of the content commons and supports authors in adapting their work. To complement this, we are currently developing a notion of "versioned references" that support the practice of creating and cultivating "islands of consistency" in the presence of change (see [KK11]). We hope that together, these measures can lead to semantic content management workflows that alleviate the side-effects of the semantic publishing workflow described in this paper.

References

[AM10] Serge Autexier and Normen Müller. "Semantics-based Change Impact Analysis for Heterogeneous Collections of Documents". In: *Proceedings of the 10th ACM symposium on Document engineering*. Ed. by Michael Gormish and Rolf Ingold. DocEng '10. Manchester, United Kingdom: ACM, 2010, pp. 97–106. ISBN: 978-1-4503-0231-9. DOI: http://doi.acm.org/10.1145/1860559.1860580. URL: http://doi.acm.org/10.1145/1860559.1860580.

[Arx] *arXMLiv Build System*. URL: http://arxivdemo.mathweb.org (visited on 09/27/2010).

[Aut+11] Serge Autexier et al. “Workflows for the Management of Change in Science, Technologies, Engineering and Mathematics”. submitted. 2011.

[Cic] *Intelligent Computer Mathematics*. submitted. 2011.

[Dav+10] Catalin David et al. “eMath 3.0: Building Blocks for a social and semantic Web for online mathematics & ELearning”. In: *1st International Workshop on Mathematics and ICT: Education, Research and Applications*. (Bucharest, Romania, Nov. 3, 2010). Ed. by Ion Mierlus-Mazilu. 2010. URL: http://kwarc.info/kohlhase/papers/malog10.pdf.

[GSK11] Deyan Ginev, Heinrich Stamerjohanns, and Michael Kohlhase. “The LaTeXML Daemon: A LaTeX Entrance to the Semantic Web”. submitted. 2011. URL: https://kwarc.eecs.iu-bremen.de/repos/arXMLiv/doc/cicm-systems11/paper.pdf.

[JK10] Constantin Jucovschi and Michael Kohlhase. “sTeXIDE: An Integrated Development Environment for sTeX Collections”. In: *Intelligent Computer Mathematics*. Ed. by Serge Autexier et al. LNAI 6167. Springer Verlag, 2010, pp. 336–344. ISBN: 3642141277. arXiv:1005.5489v1 [cs.OH].

[KK08] Andrea Kohlhase and Michael Kohlhase. “Semantic Knowledge Management for Education”. In: *Proceedings of the IEEE; Special Issue on Educational Technology* 96.6 (June 2008), pp. 970–989. URL: http://kwarc.info/kohlhase/papers/semkm4ed.pdf.

[KK11] Andrea Kohlhase and Michael Kohlhase. “Maintaining Islands of Consistency via Versioned Links”. submitted. 2011. URL: http://kwarc.info/kohlhase/submit/mkm11-verlinks.pdf.

[KMR08] Michael Kohlhase, Christine Müller, and Florian Rabe. “Notations for Living Mathematical Documents”. In: *Intelligent Computer Mathematics*. 9th International Conference, AISC, 15th Symposium, Calculemus, 7th International Conference MKM (Birmingham, UK, July 28–Aug. 1, 2008). Ed. by Serge Autexier et al. LNAI 5144. Springer Verlag, 2008, pp. 504–519. URL: http://omdoc.org/pubs/mkm08-notations.pdf.

[Koh] *General Computer Science: GenCS I/II Lecture Notes*. 2011. URL: http://gencs.kwarc.info/book/1 (visited on 03/03/2001).

[Koh+] Michael Kohlhase et al. *Planet GenCS*. URL: http://gencs.kwarc.info (visited on 09/22/2010).

[Koh06] Michael Kohlhase. OMDoc – *An open markup format for mathematical documents [Version 1.2]*. LNAI 4180. Springer Verlag, Aug. 2006. URL: http://omdoc.org/pubs/omdoc1.2.pdf.

[Koh08] Michael Kohlhase. “Using LaTeX as a Semantic Markup Format”. In: *Mathematics in Computer Science* 2.2 (2008), pp. 279–304. URL: https://svn.kwarc.info/repos/stex/doc/mcs08/stex.pdf.

[Koh+11] Michael Kohlhase et al. “The Planetary System: Web 3.0 & Active Documents for STEM”. In: accepted for publication at ICCS 2011

(Finalist at the Executable Papers Challenge). 2011. URL: https://svn.mathweb.org/repos/planetary/doc/epc11/paper.pdf.

[Plaa] *Logic Atlas and Integrator*. URL: http://logicatlas.omdoc.org (visited on 09/22/2010).

[Plab] *Planetary Developer Forum*. URL: http://trac.mathweb.org/planetary/ (visited on 01/20/2011).

[Plac] *PlanetMath.org – Math for the people, by the people*. URL: http://planetmath.org (visited on 01/06/2011).

[ZK09] Vyacheslav Zholudev and Michael Kohlhase. "TNTBase: a Versioned Storage for XML". In: *Proceedings of Balisage: The Markup Conference*. Vol. 3. Balisage Series on Markup Technologies. Mulberry Technologies, Inc., 2009. DOI: 10.4242/BalisageVol3.Zholudev01.

BauDenkMalNetz – Creating a Semantically Annotated Web Resource of Historical Buildings

Anca Dumitrache and Christoph Lange

Computer Science, Jacobs University Bremen, Germany
{a.dumitrache,ch.lange}@jacobs-university.de

Abstract. BauDenkMalNetz ("listed buildings web") deals with creating a semantically annotated website of urban historical landmarks. The annotations cover the most relevant information about the landmarks (e.g. the buildings' architects, architectural style or construction details), for the purpose of extended accessibility and smart querying. BauDenkMalNetz is based on a series of touristic books on architectural landscape. After a thorough analysis on the requirements that our website should provide, we processed these books using automated tools for text mining, which led to an ontology that allows for expressing all relevant architectural and historical information. In preparation of publishing the books on a website powered by this ontology, we analyze how well Semantic MediaWiki and the RDF-aware Drupal 7 content management system satisfy our requirements.

1 Motivation

The architectural landscape of a city is not just made up of well-established landmarks, but of historical buildings with a rich cultural background that lie outside the mainstream touristic circuit. People wanting to explore less known places of a city have little access to information about these hidden architectural gems and the stories behind them, even though all required data on historical buildings in Germany has been meticulously collected by the offices for historical monuments (Denkmalämter). However, this data has generally not been published in an easily accessible way. Existing databases and form-based search facilities are often tedious to browse through.[1]

In Bremen, an effort to collect this information and present it to the general public was made by the publisher Nils Aschenbeck, who released a series of city guide books [AW09]. However, for the moment, these books have only been published in print. By making use of these books, BauDenkMalNetz (German for "listed buildings web") proposes a way of discovering Bremen's architectural landscape that is suited for the tech-savvy tourist.

[1] See, for example, `http://194.95.254.61/denkmalpflege/index.htm`.

2 Transitioning from Written Text to Digital Media

The purpose of BauDenkMalNetz is to develop a web portal that publishes online printed text enriched with semantic annotations. Publications usually make use of a concrete set of concepts, that relate to one particular subject area, and thus can be reduced to a strict vocabulary. Identifying this vocabulary was a key step in the process of producing a formal representation of the semantic metadata that our web portal needs to store. After we have created a conceptual model of our data, we want to analyze ways of publishing our semantically enriched text online. Finally, we want to compare and contrast BauDenkMalNetz to other cultural heritage web applications, and identify possible directions for further work.

2.1 Building an Ontology

The publications that lie at the basis of our work with BauDenkMalNetz have been made available to us (but not the general public) in simple HTML files. There is a file for each individual building, with pictures associated to each file, and information like the name of the architect being highlighted. Four books have been published thus far [AW09], with more than one hundred buildings being described in total.

In order to enable enhanced browsing and querying, the data on Bremen's historical buildings needs to be organized, and the proper semantic metadata needs to be put in place. For this purpose, we have developed the BauDenkMalNetz ontology, a formal representation of the metadata vocabulary on historical buildings and related concepts, together with the relations among them. The ontology has been formalized and implemented in OWL, and was engineered in the stages specified by the METHONTOLOGY [FLGPJ97] methodology.

Scenario An example scenario of interacting with a publication backed by the BauDenkMalNetz ontology involves a tourist, working out an itinerary for visiting the city of Bremen. For this purpose, she needs to be able to browse through a particular neighborhood, by filtering the buildings based on their addresses. Suppose she is interested only in visiting those buildings that were built in the 19th century. Then she finds one particular architect that she is familiar with, and she wants to add all of his buildings to her itinerary. Finally, during her visit, she will want to stop at each individual building and read up on its history, like the years between it was built, and what famous people had been living there.

Requirements Based on this scenario, we have identified a list of requirements that the BauDenkMalNetz ontology needs to meet in order for the data to be easily accessible:

- *buildings* need to be represented as uniquely identified entities, which will be mapped to individual pages of the website; any knowledge represented using

the BauDenkMalNetz ontology needs to be interconnected, with the building entity as the central point of the representation;
- information on the *physical address* and *neighborhood* needs to be available for every building;
- the *architect* and the *architectural style* of a building have to be highlighted when that information is available;
- the *time* and *timespan* over which a building was built has to be specified for individual entries.

A more general requirement that the BauDenkMalNetz website needs to address is browsing from one building to another. This could be supported by information on the buildings' physical location (e.g. they are on the same street), or based on characteristics that they share (e.g. they were built by the same person).

Text Analysis Starting from these requirements and based on the original touristic guides, we identified they key concepts of the vocabulary that relates to historical buildings, by employing **n-gram models** [2] to find the most likely occurrences of word groupings. The results of this analysis were used in the conceptualization phase of the BauDenkMalNetz ontology. The fact that the accuracy of n-gram models increases with the volume of the processed text was an advantage that made us consider this approach.

The first step that enabled us to process the text was removing the unnecessary HTML tags, and stripping it down to a plain-text format. The text is written in German; we needed to normalize it to plain ASCII characters, as the German-specific special characters seemed to interfere with the script used to analyze it. We made use of the LaMaPUn [GJA+09] Perl library for processing the text. We used a list of the most frequent German stop words in order to filter out the information that was not meaningful for the domain vocabulary.

We analyzed series of 1 to 4-gram models. The script recognized over 600 possible groupings of words that are likely to occur together. Over 500 of these groups had a likelihood coefficient larger than 2. This coefficient is computed by having the number of incidences of the words in the group together divided by the sum of individual incidences outside of the group.

The text analysis made apparent some clear trends. Most of the likely groups of words that appeared together referred to one of the following categories: *physical buildings* (e.g. Bahnhof (*train station*) Sankt Magnus, Kirche (*church*) Sankt Magni), *personal names* (e.g. Rudolf Alexander Schroeder), *physical addresses* (e.g. Leuchtenburger Strasse (a *street*), Am Bahnhof Sankt Magnus) and *building features* (e.g. Bungalow, Turm (*tower*)). By identifying these categories, we got a first impression of what are the key concepts we need to define for our ontology.

[2] A probabilistic model that, given the first $n-1$ words in a sentence, will predict the n^{th} word. [MS99]

Conceptualization Based on this analysis, and according to the requirements identified in the previous section, we conceptualized entities to be represented in the BauDenkMalNetz ontology[3]. Most concepts identified during the n-gram analysis were transformed into resources, then properties were added to connect them. The core of the BauDenkMalNetz ontology is the following (concepts underlined, relations in *italics*):

- building – a resource identifying a particular building;
- building part – a subconcept of the building entity (e.g. tower, annex);
- building complex – a composite consisting of several building entities;
- building type – different types of constructions (e.g. church, hospital);
- *address* – the physical location of a building;
- *architect* – the person or group of people that have designed the building;
- *inhabitant* – famous person that has lived in that building;
- year – *when a building was built*; can refer to the year when *construction began*, *ended*, or both.

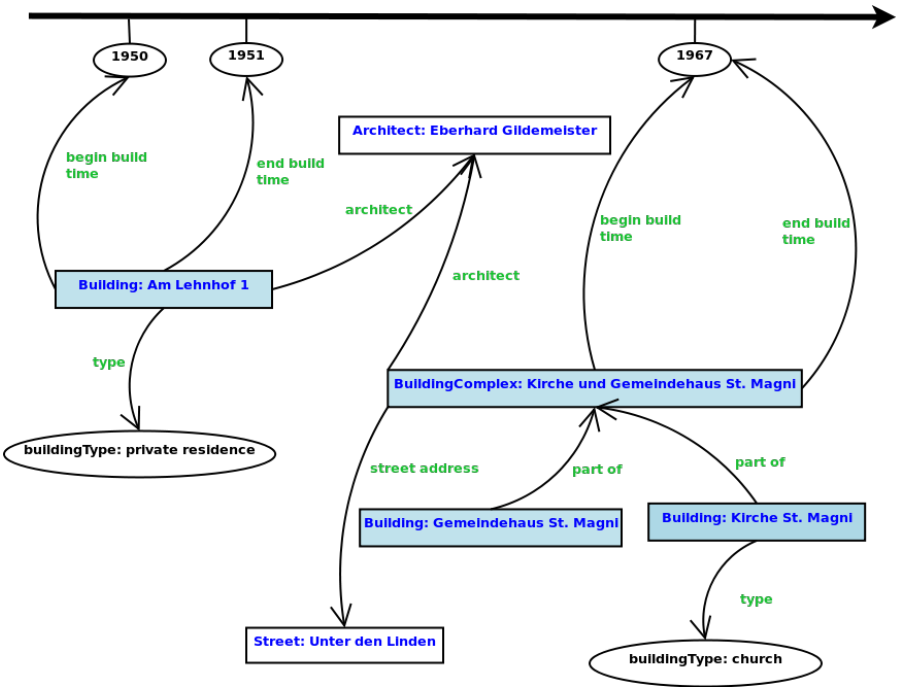

Fig. 1. A fragment of the BauDenkMalNetz ontology

[3] Available at: `http://oaff.info/ontology/bdmn#`.

Alignment to Other Ontologies The Linked Data community [Hea+] advocates the reuse of knowledge models and vocabularies, in order to achieve interoperability across the Web. Indeed, there already exist various ontologies that model some of the relevant knowledge about historical buildings, out of which we found the following ones relevant for aligning with the BauDenkMalNetz ontology:

- The **GeoNames** [Geo] ontology models geospatial semantic information. In particular, it assigns to individual locations on the globe a unique URI. For our purposes, it can be used to uniquely identify each historical building based on its coordinates. Reusing this ontology brings the added advantage of explicitly specifying the geolocation of a building, which allows for easier integration with web mapping services.
- The **CIDOC CRM** [Cid] ontology represents the detailed scientific documentation of cultural heritage objects, which include historical monuments. By aligning our ontology to CIDOC CRM, we can formulate a full description of the historical information related to a building (e.g. the architectural style of the monument, the official sources which document the monument etc.).

2.2 Publishing in a Semantic Content Management System

For deploying BauDenkMalNetz, we have so far established requirements and analyzed how well two semantic content management systems satisfy these requirements: **Semantic MediaWiki** (SMW [Sem]) and **Drupal 7** [Dru].

Requirements Based on the scenario discussed in the previous section, we have also analyzed the requirements that our website needs to provide. Digitally representing publications means that the BauDenkMalNetz web portal needs to build on the use cases of the written text that lies at its core, and enhance them with semantic browsing and querying capabilities that will provide for a better user experience. Therefore, a suitable content management system for deploying BauDenkMalNetz should offer the following functionality:

1. the possibility of integrating RDF triples, and at least a minimum of ontology support;
2. support for querying the RDF content of the website (e.g. by using SPARQL);
3. browsing based on the semantic metadata;
4. extensible publishing support for:
   (a) users, through enabling PDF and HTML exporting;
   (b) machines, by interlinking the publications across the Web, according to linked data principles;
5. the possibility of importing large amounts of text into the system.

Semantic MediaWiki SMW [Sem] was built as an extension of MediaWiki, the wiki engine which powers Wikipedia. It provides enhanced features for browsing and organizing its contents via semantic annotations. We built the first BauDenkMalNetz prototype using SMW [DLK+10].

Our motivation for using SMW in deploying the initial version of our web portal was its suitability for rapidly creating a working prototype (cf. [BDH+09]). SMW allows for easily adding and editing of the necessary data and metadata available on historical buildings, in keeping with requirements 1 and 3. New information could be easily incorporated and linked to the already existing data via SMW's page creation and editing tools. At the same time, the metadata vocabulary (i.e. the ontology) could be easily modified, simply by adding in-text annotations.

Requirement 2 is addressed by a simple query language included in SMW. The SMW querying functionality does not operate directly on RDF, and instead uses a syntax that addresses RDF triples based on the names with which they are declared in the wiki pages. While it provides basic functionality for querying RDF data, which includes selecting pages in the wiki, together with what properties of the pages to display, the SMW query language lacks the complexity of SPARQL (e.g. querying within a particular namespace).

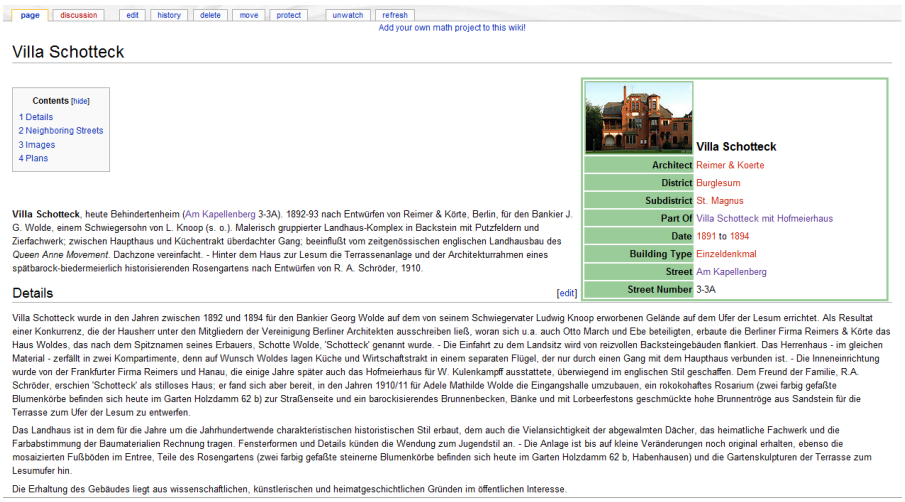

Fig. 2. Screenshot of the SMW prototype

When further assessing requirement 1, we found that the conceptual model of our metadata was less obvious and never explicitly formalized, as the ontology, to which the texts adhere, is not necessarily specified explicitly in SMW, but rather implied from the annotations done directly on the text. In this case, alignment to other similar ontologies (in keeping with the linked-data philosophy of reuse)

is still possible, yet it is rendered more difficult by the lack of an explicit formal definition of the ontology.

Requirement 5 was also not addressed by our prototype. SMW provides some tools suited for database import, however the texts we want to analyze are stored in simple HTML files. The volume of data that needs to be processed makes it almost impossible to have the texts annotated manually, like we did for building the prototype, while also making BauDenkMalNetz rather suited for the employment of natural language processing techniques in order to get the needed semantical annotations.

Drupal 7 As our goal is to publish existing content, rather than creating new content in a collaborative way, we also considered **Drupal** [Dru], a rather traditional content management system. Given the BauDenkMalNetz documents collection and our ontology, we have so far analyzed Drupal's features w.r.t. the requirements established above. Deploying BauDenkMalNetz in Drupal remains to be done in spring 2011.

Requirement 1 is satisfied as the latest version 7 of Drupal provides an RDF API [CDC+09] that is integrated in the Drupal core. This enabled us to easily upload our OWL ontology into the website, by using the RDF vocabulary import feature. The keywords pertaining to each resource were then added to the taxonomy of our website, and mapped to the corresponding classes and properties in the ontology. For printed media, where a particular text usually does not undergo much change after being published, the advantage that Drupal brings is that, as the structure of the text is already known, its conceptualization can be set as the core of the website via the RDF API even before the website is deployed.

Requirement 2 is addressed by the SPARQL module for Drupal, which allows us to query our external triple store. The task of building meaningful queries is made even easier by the SPARQL Views [Cla10] module, which supports visual query building and result display.

Results When comparing SMW to Drupal, we have encountered some drawbacks of SMW that led us to reconsider our approach. The flexibility and agility of SMW were not of a particular advantage in our setting. The publication sources are imported from external sources, and therefore we are not interested in MediaWiki's collaboration support. The ontology and its connections to other ontologies are, for now, created just by us, but they are not evolved or extended dynamically by a community – therefore we are not interested in giving write access to the ontology via the content management system. We rather prefer having a clear conceptual model of the metadata from the beginning. Drupal supports the initial import of such an ontology before importing the content and thus is suited for managing annotations to publications that have already existed before.

Also, we have concluded that using SPARQL to power our query engine would provide more flexibility for our queries, while also making them portable,

as SPARQL is not platform dependent. While SMW is currently working to integrate SPARQL[4] functionality in its core, for the moment, the support it provides is limited, whereas Drupal provides SPARQL support through the modules discussed in the previous section.

Table 1. Comparison of SMW and Drupal based on the requirements list.

Req.	SMW	Drupal	Results
1.	inline RDF triples declaration, no explicit ontology support	RDF part of the core, Evoc module for ontology import	*Drupal* for better ontology support
2.	SMW query language	SPARQL, SPARQL Views modules	*Drupal* for advanced querying possibilities
3.	wiki pages mapped to resources and categories	RDF mapping for content types	draw
4a.	third-party plugin, not well documented	Printer, e-mail and PDF versions module in development[a]	*Drupal*
4b.	synchronizing with vocabularies supported by SMW through export[b] and import[c]	Evoc external vocabulary support	draw
5	through page creation, with manual semantic annotations	through page creation, but with specialized content types	*Drupal*

[a] `http://drupal.org/project/print`
[b] `http://semantic-mediawiki.org/wiki/Help:RDF_export`
[c] `http://semantic-mediawiki.org/wiki/Help:Import_vocabulary`

3 Development and Evaluation Plan

During spring 2011, we continued developing the BauDenkMalNetz website in Drupal, by uploading the texts of the tourist guides to our website, with the keywords in the vocabulary highlighted in the resource's pages. We will make semantic browsing available, based on these key concepts, achieved through Drupal's taxonomy feature. Also, for increased functionality, we will add a geospatial aspect to the semantic navigation by utilizing the Google Maps

[4] `http://semantic-mediawiki.org/wiki/SPARQL_and_RDF_stores_for_SMW`

API [Goo]. Finally, resources referring to people (e.g. the architect) will be cross-referenced with Linked Data resources, like DBPedia[5].

For even more advanced querying features, we are considering to make use of the **XSPARQL** [AKK+08] query language. XSPARQL combines the XML query language XQuery with the RDF query language SPARQL, which allows for generating XML-formatted results for queries over the semantic metadata of our website and, in future, interlinked websites. By selecting from a list of available queries, tourists will be able to create personalized guides of historical buildings.

For evaluating the usability of the BauDenkMalNetz website, existing methods for evaluating (semantic) digital libraries [FTA+07; Kru09] are applicable. A group of test-users will navigate through the website, providing feedback based on *usability* (of the content management system with our extensions) and *usefulness* (of the content, in the way our system publishes it). The users will provide feedback on how easy/difficult it is to find a particular building, by querying the system based on a criteria of their own choosing (e.g. location, architectural style etc.), and also about how they managed to find their way from one particular building to another, based on a common characteristic. They will also be asked to provide their input on how accurate the query results are in relation to what they were expecting to find, and also about the informative character of individual buildings' pages. Based on this assessment the user-friendliness of the website we will consider possible improvements. A first release of BauDenkMalNetz, adapted according to the results of an initial evaluation round, is expected in May.

4 Related Work on Cultural Heritage

There exist a number of projects that process data about cultural heritage using semantic web technologies. Most approaches encountered gather the information from a wide array of sources (e.g. historical documents, archaeological excavation reports etc.), and consequently one of their main issues is developing an ontology that serves as a common medium for these different types of texts. In contrast, the BauDenkMalNetz ontology was developed from a singular source – published texts written in the same style, by the same author, on the topic of cultural heritage. Therefore, the ontology's intended use is not to provide a universal definition of the vocabulary describing historical buildings, but to define the vocabulary used by this particular series of publications. By studying the related work on cultural heritage we were able to shed some light on how we could improve our data model in order to represent a greater pool of sources, therefore enabling the reusability of our core ontology. For this purpose, the following applications have been assessed:

MANTIC [MPV10] is a project similar to BauDenkMalNetz, that represents data on cultural heritage sites of the city of Milan, that was gathered from historical sources and publications. At its core, it uses the CIDOC CRM ontology for storing information about the archeology of the city. This information is then

[5] `http://dbpedia.org`

incorporated into the Google Maps API, making for an easy to use application for browsing Milan's historical landmarks, that is quite similar in scope to our work. Unlike BauDenkMalNetz, MANTIC deals with historical sources, which comprise a great variety of publications, written in different styles and over a long period of time. MANTIC provides a good example of how CIDOC CRM can be reused for representing historical landmarks, however, since the sources MANTIC deals with are so disjointed, identifying a common vocabulary for them is more difficult, and therefore no special ontology that deals primarily with historical buildings was devised.

The **Fundación Marcelino Botín** [Fun] worked on a similar project that aimed to gather information on eleven cultural heritage sites of Cantabria, a region of Northern Spain. Like MANTIC, the Cantabria project had to reconcile information from a heterogeneous set of sources, by adapting the CIDOC CRM ontology to suit their dataset. However, most of the data populating the ontology had already been preprocessed (as spreadsheets, web pages etc.), and adding content to the project website was done in a semi-automated way. Therefore, unlike BauDenkMalNetz, the Cantabria project is intended as a community portal, where experienced users can modify or add new data to the website and to the ontology. Aside from providing another example of how to reuse existing standards, this project is relevant for us because of the way it makes use of the various benefits brought by using semantic metadata: a semantic search engine, an interactive map based on geoposition metadata, and interoperability with other cultural heritage repositories.

CultureSampo [HMK+09] is an application that publishes cultural heritage information about Finland. Like BauDenkMalNetz, CultureSampo builds on existing standards for conceptualizing cultural items, and then extends them with domain specific information. However, as it covers a larger content (history, folklore, artifacts etc.), CultureSampo integrates a wide array of domain specific ontologies, that were developed in a semi-automatic fashion based on existing thesauri. While the development methodology of CultureSampo is relevant and can be adapted for BauDenkMalNetz, the scope of the project is too wide to enable us to reuse their data model.

5 Conclusion and Further Work

After assessing in which ways traditional printed publications on historical landmarks can be enhanced by transposing them in a digital format and enriched with semantic annotations, we devised the BauDenkMalNetz ontology, by analyzing its requirements and processing the texts that were made available to us by using text mining techniques. In keeping with linked data principles, we aligned our ontology to other existing representations that relate to our specific domain, like CIDOC CRM and GeoNames. Once we determined the structure of our metadata, we compared how different content management systems (SMW and Drupal 7) satisfy the requirements for deploying the BauDenkMalNetz website. As Drupal provides a more rigorous way of declaring a conceptual model, which is more

suitable for digital publications, we have chosen it as the medium in which our web portal will be developed.

Once finished, the BauDenkMalNetz website will provide a comprehensive and easy-to-use guide to the city of Bremen, and possibly even help boost the touristic appeal of Bremen. A possible enhancement to the resource will be creating a mobile version of the website, so that tourists can create virtual itineraries that they can access on the go. However, the scope of our work is not limited to Bremen. We believe that both the ontology and the vocabulary will prove general enough to adapt in order to represent any touristic publication guide on historical landmarks.

Acknowledgments

The authors would like to thank Deyan Ginev for help with the LaMaPUn library, Lin Clark for help with assessing Drupal 7, and the anonymous peer reviewers for their pointers to further related work.

References

[AKK+08] W. Akhtar, J. Kopecký, T. Krennwallner, et al. "XSPARQL: Traveling between the XML and RDF worlds – and avoiding the XSLT pilgrimage". In: *The Semantic Web: Research and Applications*. 5th European Semantic Web Conference (ESWC) (Tenerife, Spain). Ed. by S. Bechhofer, M. Hauswirth, J. Hoffmann, et al. LNCS 5021. Springer Verlag, 2008.

[AW09] N. Aschenbeck and I. Windhoff. *Landhäuser und Villen in Bremen*. Bremen: Aschenbeck Verlag, 2009.

[BDH+09] J. Bao, L. Ding, R. Huang, et al. "A Semantic Wiki based Light-Weight Web Application Model". In: *Proceedings of the 4th Asian Semantic Web Conference*. 2009, pp. 168–183.

[CDC+09] S. Corlosquet, R. Delbru, T. Clark, et al. "Produce and Consume Linked Data with Drupal!" In: *The Semantic Web*. 8th International Semantic Web Conference (ISWC). Ed. by A. Bernstein, D. R. Karger, T. Heath, et al. LNCS 5823. Springer, Oct. 2009.

[Cid] *The CIDOC Conceptual Reference Model*. URL: `http://cidoc.ics.forth.gr` (visited on 2010-03-07).

[Cla10] L. Clark. "SPARQL Views: A Visual SPARQL Query Builder for Drupal". In: *Poster and Demo Proceedings of the 9th International Semantic Web Conference (ISWC)*. 2010. URL: `http://iswc2010.semanticweb.org/pdf/518.pdf`.

[DLK+10] A. Dumitrache, C. Lange, M. Kohlhase, et al. "Prototyping a Browser for a Listed Buildings Database with Semantic MediaWiki". In: *5th Workshop on Semantic Wikis*. Ed. by C. Lange, J. Reutelshöfer, S. Schaffert, et al. CEUR Workshop Proceedings 632. 2010. URL: `http://ceur-ws.org/Vol-632/`.

[Dru] *Drupal.org – Community plumbing*. web page at `http://drupal.org`. URL: `http://drupal.org`.

[FLGPJ97] M. Fernández-López, A. Gómez-Pérez, and N. Juristo. "METHONTOLOGY: from Ontological Art towards Ontological Engineering". In: *Proceedings of the Fourteenth National Conference on Artificial Intelligence AAAI-97*. (Stanford, USA). MIT Press, 1997, pp. 33–40.

[FTA+07] N. Fuhr, G. Tsakonas, T. Aalberg, et al. "Evaluation of digital libraries". In: *International Journal of Digital Libraries* 8 (2007), pp. 21–38.

[Fun] *Case Study: An Ontology of Cantabria's Cultural Heritage*. URL: `http://www.w3.org/2001/sw/sweo/public/UseCases/FoundationBotin/` (visited on 2011-04-12).

[Geo] *GeoNames*. URL: `http://www.geonames.org` (visited on 2010-04-23).

[GJA+09] D. Ginev, C. Jucovschi, S. Anca, et al. "An Architecture for Linguistic and Semantic Analysis on the arXMLiv Corpus". In: *Applications of Semantic Technologies (AST) Workshop, Informatik*. 2009. URL: `http://www.kwarc.info/projects/lamapun/pubs/AST09_LaMaPUn+appendix.pdf`.

[Goo] *Google Maps*. URL: `http://maps.google.com` (visited on 2011-01-10).

[Hea+] T. Heath et al. *Linked Data – Connect Distributed Data across the Web*. URL: `http://linkeddata.org` (visited on 2010-06-11).

[HMK+09] E. Hyvönen, E. Mäkelä, T. Kauppinen, et al. "CULTURE SAMPO – A National Publication System of Cultural Heritage on the Semantic Web 2.0". In: *ESWC*. 6th European Semantic Web Conference (ESWC). Ed. by L. Aroyo, P. Traverso, F. Ciravegna, et al. LNCS 5554. Springer, 2009.

[Kru09] S. R. Kruk. "Semantic Digital Libraries. Improving Usability of Information Discovery with Semantic and Social Services". PhD thesis. National University of Ireland, Galway, 2009.

[MPV10] G. Mantegari, M. Palmonari, and G. Vizzari. "Rapid Prototyping a Semantic Web Application for Cultural Heritage: The Case of MANTIC". In: *The Semantic Web: Research and Applications (Part II)*. 7th Extended Semantic Web Conference (ESWC). Ed. by L. Aroyo, G. Antoniou, E. Hyvönen, et al. LNCS 6089. Springer, 2010.

[MS99] C. D. Manning and H. Schütze. "Statistical Inference: n-gram Models over Sparse Data". In: *Foundations of Statistical Natural Language Processing*. MIT Press, 1999. Chap. 6.

[Sem] *Semantic MediaWiki*. URL: `http://semantic-mediawiki.org` (visited on 2010-03-04).

Sustainability of Evaluations Presented in Research Publications

Raúl García-Castro

Ontology Engineering Group,
Departamento de Lenguajes y Sistemas Informáticos e Ingeniería Software
Facultad de Informática, Universidad Politécnica de Madrid, Spain
rgarcia@fi.upm.es

Abstract. This position paper discusses how research publication would benefit of an infrastructure for evaluation entities that could be used to support documenting research efforts (e.g., in papers or blogs), analysing these efforts, and building upon them. As a concrete example in the domain of semantic technologies, the paper presents the SEALS Platform and discusses how such platform can promote research publication.

1 Introduction

The way of publishing evaluation-related information in research papers should be rethought to facilitate the use of such information.

The limited extension of research papers does not allow a full description of all the entities involved in evaluations (evaluation workflows, test data, tools and results) and of the concrete context in which evaluations were performed. Therefore, it is difficult if not impossible to reproduce evaluations described in research papers and to validate them; this forces researchers in most cases to blindly trust in the paper claims. Besides, both technologies and evaluation methods evolve over time and the evaluation data included in the paper becomes rapidly outdated.

Another perspective to take into account is that of advancing research by building upon existing one. Defining and performing evaluations is expensive and prone to errors. This is mainly because, besides the lack of full evaluation descriptions mentioned above, most lessons learnt during evaluation (both positive and negative) are not explicit in research papers. Furthermore, performing complex analyses across research papers (e.g., finding correlations between the results of different evaluations) is currently not possible.

The goal of this paper is to discuss how research would benefit of an infrastructure for evaluation entities that could be used to support documenting research efforts (e.g., in papers or blogs), analysing these efforts, and building upon them.

Such infrastructure would allow anyone reading or reviewing a paper to completely analyse the evaluations presented in the paper and to validate them, taking advantage of dynamic and enhanced result visualisations.

Besides, it would permit anyone to reproduce the evaluation presented in the paper under the same settings or using updated or alternative versions of the evaluation entities.

Furthermore, anyone interested in building upon existing evaluations could reuse the evaluation presented in the paper (fully or parts of it) and even combine the results from evaluations in different papers.

Clearly, someone could disagree with the above-mentioned claims; the main stands against them could be the following:

- *Refusal to unveil evaluation details.* In research environments, people are used to having other people review their work in detail, so this should be no problem. Besides, being against this would incline people to think that the researcher is hiding something.
- *Refusal to share work with others.* This opinion is also not expected since in research environments people are usually eager to be reused cited.
- *Refusal to devote effort to share evaluation details.* Even if researchers acknowledge the added value of sharing their evaluations, they will be reluctant to do so unless the benefits compensate their spent efforts.
- *Refusal to reuse work from others.* Even if the do-it-yourself attitude is characteristic of computer science researchers, they are also aware of the benefits of reuse; therefore, this is something that should not pose rejection.

The SEALS (Semantic Evaluation at Large Scale) European project[1] is developing an infrastructure (the SEALS Platform) that offers independent computational and data resources for the evaluation of semantic technologies [1].

Next, the paper presents an overview of the SEALS Platform and then discusses how such infrastructure could support the publishing and management of evaluation information in research papers, providing different benefits along the lines presented above.

2 An Infrastructure for Semantic Technology Evaluation

The idea of software evaluation followed in the SEALS Platform is largely inspired by the notion of evaluation as defined by the ISO/IEC 14598 standard on software product evaluation [2]. In any *evaluation* a given set of *tools* are executed, following a given *evaluation workflow* and using determined *test data*. As an outcome of this process, a set of *evaluation results* is produced.

This high-level classification of software evaluation entities can be further refined as needed; a detailed description of them and their life cycles can be found in [3]. For example, in accordance with the approach followed in the IEEE 1061 standard for a software quality metrics methodology [4], evaluation results are classified according to their provenance, differentiating *raw results* (those evaluation results directly generated by tools) from *interpreted results* (those generated from other evaluation results).

[1] `http://www.seals-project.eu/`

Moreover, our entities include not only the results obtained in the evaluation but also any contextual information related to such evaluation, a need also acknowledged by other authors [5]. To this end, we also represent the information required for automating the execution of an evaluation description in the platform, which, with the rest of the entities presented, yields traceable and reproducible evaluation results.

The SEALS Platform has been developed around these evaluation entities and following a service-oriented approach. The architecture of the platform comprises a number of components, shown in Figure 1, which are described below.

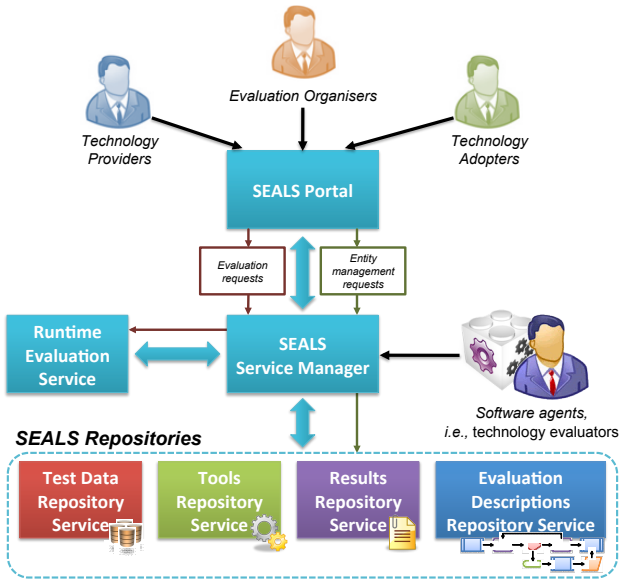

Fig. 1: Architecture of the SEALS Platform.

- **SEALS Portal.** The SEALS Portal provides a web user interface for interacting with the SEALS Platform. Thus, the portal will be used by the users for the management of the entities in the SEALS Platform, as well as for requesting the execution of evaluations.
- **SEALS Service Manager.** The SEALS Service Manager is the core module of the platform and is responsible for coordinating the other platform components and for maintaining consistency within the platform. This component exposes a series of services that provide programmatic interfaces for the SEALS Platform. Thus, apart from the SEALS Portal, the services offered may be also used by third party software agents.
- **SEALS Repositories.** These repositories manage the entities used in the platform: test data, tools, results, and evaluation workflows.

– **Runtime Evaluation Service.** The Runtime Evaluation Service is used to automatically evaluate a certain tool according to a particular evaluation description and using some specific test data.

All the evaluation entities stored in the platform are described according to a set of OWL ontologies[2] [3]. Since the entities presented above share a number of common properties, we developed an upper ontology to represent them, as well as different ontologies covering each entity domain. During the definition of the ontologies we tried, when possible, to reuse current standards and models (i.e., Dublin Core, FOAF, VCard).

3 Publication and Management of Evaluation Information

This section discusses how the SEALS Platform could support the publication and management of evaluation information.

The SEALS Platform offers manual and programmatic access to the evaluation entities stored in its repositories. This allows linking the evaluation resources mentioned in research papers to the actual resources stored in the platform. Besides, if reverse links were created from the evaluation entities to research papers, networks of papers around concrete evaluations could be built.

The SEALS Platform also allows storing different versions of tools, evaluation workflows and test data. This way, it maintains the traceability from the concrete evaluation used in one paper to those evaluations that include updated versions of tools, evaluation workflows or test data.

All the evaluation entities stored in the SEALS Platform are described using ontologies with the aim of having consensual and interoperable descriptions. These machine-processable descriptions can be published in the Web or be embedded in research papers. Furthermore, it provides dynamic and interactive visualisations of evaluation results that could be used in non-standard research papers (e.g., multimedia or interactive documents).

In the SEALS Platform, evaluation reproducibility is a main requirement. To this end, evaluations are only executed over persistent (i.e., unmodifiable) entities and the whole evaluation execution context is stored. This allows replicating the concrete evaluation presented in a research paper at any moment and by anyone.

All the evaluation entities can not only be accessed but also be reused both inside and outside the SEALS Platform. This reuse can be performed as a whole (e.g., reusing some test data in another evaluation infrastructure) or partially (e.g., evaluation workflows are defined with the BPEL language and new workflows can be defined from existing workflows and services).

Finally, since evaluation results are represented following common schemas (i.e., ontologies), researchers could exploit these results in unexpected ways. To allow this, we have defined a quality model for semantic technologies that defines the main quality characteristics of such technologies and allows the combination and comparison of results from different evaluations [6].

[2] `http://www.seals-project.eu/ontologies/`

4 Conclusions

This paper proposes to support research publications (or any other type of research documentation) through an infrastructure for evaluation entities. Having such infrastructure would allow, on the one hand, connecting research publications with the actual evaluations used in them and, on the other hand, interconnecting different research efforts.

The SEALS Platform aims to support these ideas in the domain of semantic technologies with the ultimate goal of increasing the maturity of the semantic research community by enriching the body of knowledge on semantic technology evaluation and by encouraging an experimentation-based research.

However, the project is still in its way to achieve the approach presented in this paper since functionalities for linking evaluations with publications are not planned yet. To this end, future challenges to be faced are not only technological but also social (e.g., it requires greater commitment since researchers have to invest more effort than they are now) or legal (e.g., important issues are the access and use policies for evaluation data).

Furthermore, the success of such approach will depend on the existence of software technologies that are coupled to researchers' working environments and that leverage the effort of using an infrastructure such as the SEALS Platform in day-to-day research.

Acknowledgements

This work has been supported by the SEALS European project (FP7-238975).

References

1. García-Castro, R., Esteban-Gutiérrez, M., Gómez-Pérez, A.: Towards an infrastructure for the evaluation of semantic technologies. In: Proceedings of the eChallenges 2010 Conference, Warsaw, Poland (2010)
2. ISO/IEC: ISO/IEC 14598-6: Software product evaluation - Part 6: Documentation of evaluation modules (2001)
3. García-Castro, R., Esteban-Gutiérrez, M., Kerrigan, M., Grimm, S.: An ontology model to support the automatic evaluation of software. In: Proceedings of the 22nd International Conference on Software Engineering and Knowledge Engineering (SEKE 2010), Redwood City, CA, USA (2010) 129–134
4. IEEE: IEEE 1061-1998. IEEE Standard for a Software Quality Metrics Methodology (1998)
5. Kitchenham, B.A., Hughes, R.T., Linkman, S.G.: Modeling software measurement data. IEEE Trans. Softw. Eng. **27** (2001) 788–804
6. Radulovic, F., García-Castro, R.: Towards a Quality Model for Semantic Technologies. In: Proceedings of the 11th International Conference on Computational Science and Its Applications (ICCSA 2011), Santander, Spain, (To be published) (2011)

A semantic model for scholarly electronic publishing

Carlos H. Marcondes

University Federal Fluminense, Department of Information Science, R. Lara Vilela, 126, 24210-590, Niterói, Rio de Janeiro, Brazil, marcon@vm.uff.br

Abstract. Despite numerous advancements in information technology, electronic publishing is still based on the print text model. The natural language textual format prevents programs from semantically processing article content. A semantic model for scholarly electronic publishing is proposed, in which the article conclusion is specified by the author and recorded in a machine-understandable format, enabling semantic retrieval and identification of traces of scientific discoveries and knowledge misunderstandings. 89 biomedical articles were analyzed for this purpose. A prototype system that partially implements the proposed model was developed. Four patterns of reasoning and sequencing of semantic elements were identified in the analyzed articles. A content model comprising semantic elements and their sequences in articles is proposed. The development and testing of a prototype of a Web submission interface to an electronic journal system that implements the proposed model are reported.

Key words: electronic publishing, scientific methodology, scientific communication, knowledge representation, ontologies, semantic content processing, e-Science

1 Introduction

Before the advent of the World Wide Web (hereafter referred to as "Web"), man's body of scientific knowledge was fuzzy and distributed across publications in libraries worldwide. The Web is fast becoming a universal platform for the disposal, exchange, and access of knowledge records. An increasing amount of records of human culture—from text, static and motion images, and sound, to multimedia—are now being created directly in a digital format.

With regard to scientific knowledge, one problem is the fact that although a large amount of knowledge can potentially be made available through the Web in digital formats, this knowledge is embedded in the text of scientific articles in natural language that is only comprehensible to humans. Scholarly electronic publishing is based on the print text model. These texts are also distributed across various information resources such as digital libraries, electronic journal systems, and repositories. Their textual format hinders the comparison of their semantic content by computers in order to identify gaps and contradictions and agreements in knowledge.

Metadata is essential for managing knowledge records in an increasingly complex digital environment. Since the MARC (machine-readable cataloging) record was established in the 1960s, bibliographic record models have hardly changed. A typical bibliographic record comprises sets of database fields, including a flat space of a list of unconnected fields for content description, where keywords or descriptors are assigned, each having an equal weight for retrieval purposes. Content access to documents in modern bibliographic retrieval systems is still achieved by matching user queries formed by keywords connected by Boolean operators to keywords comprising the bibliographic records, in a manner similar to early bibliographic retrieval and library automation systems.

A subtle distinction, rarely made by the Library and Information Science Community, must be made between the *aboutness* of a document, a concept that has been exhaustively discussed in this community, and the *claims* made by authors throughout the text of the documents. Indexing activities address the former but not the latter. The extraction and representation in machine-understandable format of claims in scientific article texts should constitute a step toward conventional information retrieval (IR) systems. It should enable direct knowledge management, its use in automatic reasoning and inference tasks applied to different and unpredicted contexts, and increased possibilities of the automatic processing of the rich digital content now available throughout the Web.

Relations between concepts are the core of meaning. Dictionary entries with definitions of terms, thesauri, and classification schemas are examples of this claim. Typical bibliographic records do not hold explicit semantic relations between elements comprising the content of documents they represent. Boolean operators are too general and lack the semantic expressiveness necessary for content retrieval in

specific scientific domains. Relations expressed by Boolean operators are processed as extensive set operations on the keywords included in the bibliographic records, and not as intensive semantic relations.

In comparison with the poor expressiveness of the three Boolean operators, the UMLS (unified medical language system) Semantic Network (hereafter abbreviated as "SN") [1], which is the classification schema of the UMLS NIH (National Institutes of Health) Metathesaurus, organizes every concept in hierarchy trees, each having as its root a top level Semantic Type. The UMLS SN uses 54 Relation Types to express the semantic relations used between concepts in Semantic Type hierarchies used to index Biomedical Science scientific articles. The UMLS SN holds the permitted relations between Semantic Types. Although this semantically richer schema is supported by the UMLS, the bibliographic record models in databases such as Medline are incapable of exploiting this potential.

Semantic Web (SW) technologies [2] constitute a step toward semantic retrieval and processing in computational environments. The proposal content of a Web document is no longer a matter of keyword match as in conventional computational environments since the 1960s, but instead comprises structured sets of concepts connected by precise meaning relations as in RDF (Resource Description Framework) [3] and RDF Schema [4] statements. Such a rich knowledge representation schema enables software agents to perform "inferences" and more sophisticated tasks based on the document content.

Since the Actas of the Royal Society in the seventeenth century, scientific articles have become privileged channels of scientific communication. Through scientific articles, authors bring discoveries into the public knowledge. Nowadays, scholars and researchers commonly engage in electronic Web publishing. Most scientific journals are now available on the Web. Modern bibliographic information systems exploit the potential of information technology (IT). However, IT is not yet used to directly process the knowledge embedded in the text of scientific articles. Electronic-Web-published articles can serve as *knowledge bases,* as stressed by Gardin [5]. However, in the digital format, these knowledge bases are useful only to humans, who can read them. The content of scientific articles deserves critical reading, inquiry, and citation through a long social process until it becomes part of man's body of knowledge.

In the present proposal, a richer semantic content bibliographic record model is proposed, in which scientific claims made by authors throughout articles are expressed by relations between phenomena. In the proposed model, each article, in addition to being published in textual format, has its claims also represented as structured relations and recorded in a machine-understandable format using SW standards such as RDF [3] and OWL (Web Ontology Language) [6]. In the proposed model, article records comprise full-text, conventional bibliographic metadata, and semantic metadata conveying the claims made by the author. The machine-understandable records resulting from this publishing model can be compared by software agents either with public knowledge—e.g., published scientific articles—or with terminological knowledge bases throughout the Web, thus providing scientists with new tools for knowledge retrieval, claim comparison, identification of contradictory claims, use of these claims in different contexts, and identification and validation of new contributions to science made by specific articles.

We propose to engage authors in developing a richer content representation of their own articles; bibliographic record instances in compliance with the proposed model will be generated by a Web author's submission interface to a journal system, as a byproduct of submitting his/her articles to the system. Such a system, during the upload process of scientific article files, will perform an interactive dialog with authors in order to extract the semantic content of the claims made in the scientific articles and record them in a machine-readable format. We also report the initial steps toward the development of such a system.

Several alternatives have already been proposed as new types of publications that address the previously discussed issues; to try and exploit SW technologies to enhance scientific communication, management, sharing, and reuse of knowledge; and to provide direct access to semantic content of scientific articles. Thus, there is an increasing trend in electronic publishing experiences toward formalizing the text of articles or structuring them, marking them, and identifying significant parts to facilitate more direct reading by humans, potentially by relating the text to formal ontologies [7] as a means to overcome the ambiguity of the texts and allow their "semantic" processing by programs.

The remainder of this article is organized as follows. The next section presents a review of the theoretical concepts the proposed publication model is based on along with similar experiences and projects. Section 3 describes the materials and methods used. Section 4 describes the model, its elements, and the development of a prototype system of a Web author's submission interface to a journal system, which partially implements the model. Finally, section 5 presents the results obtained thus far and discusses the conclusions. It also outlines the future research steps.

2 Related studies

From an ontological point of view, scientific articles are (a) documents embedded in definite social relations concerning the scientific communication protocols exhaustively studied in Information Science [8], [9], and with regard to their textual structure, (b) a text-embedded rhetoric/logical theory [10], [11], [12]. The focus of the proposed model is the second aspect, i.e., the reasoning/rhetorical, and the semantic structure of the scientific articles in Biomedical Sciences.

In this field in particular, new research methods challenge the conventional Scientific Method and Popperian hypothesis-driven research. The so-called high-throughput methods like DNA microarrays and proteomics [13] allow scientists to process a great amount of data rapidly and in parallel, thus "conducting experiments about which no predictions can be made because no hypotheses have been constructed," as stressed by Westein [14]. This author also stresses the following:

> "Given the layered, evolutionary complexity of biological systems, it will not be possible to understand them comprehensively on the basis of hypothesis-driven research alone. Likewise, it will not be possible to do so solely through "omic" studies of genes, proteins, and other molecules in aggregate. The two modes of research are complementary and synergistic".

Several alternatives have been considered as new types of publications to address the previously discussed issues and to exploit SW technologies to enhance scientific communication, management, sharing, and reuse of knowledge, and to provide direct access to the semantic content of scientific articles. The following text comments on these experiences and their conceptual bases.

The Prospect project is a publishing initiative of the Royal Society of Chemistry, in which terms in the texts of articles that refer to chemical or biological entities have links to dictionaries or ontologies that define them. The Elsevier publishing group is developing a project called Article of the Future associated with the biomedical journal Cell in order to add functionality to several articles, including change in presentation (hierarchical presentations), summary charts, and a section on "Highlights" that briefly outline the conclusions of the article. These facilities are only possible in a Web environment for digitally published articles. Sample articles are available on the project Web site to demonstrate these facilities. A previous study [15] has described the experience of using different semantic technologies in the journal PloS, including biomedical ontologies, comments on the articles, and an ontology of types or reasons for citation.

HyBrow [16] is a system aimed at helping scientists with hypothesis formulation and evaluation against previous knowledge. The work by Hunter and co-authors [17] aimed to identify concepts for extracting protein interaction relations from biomedical text. The approach of [18] to semantic annotations in medical articles considers assertions to be the fundamental units of knowledge. The HypER approach [19] also considers claims to be the basic unit of scientific knowledge. Groth and colleagues [20] present a publication model called nanopublications, consisting of core scientific statements associated with their annotations which specify their context; scientific statements are coded as RDF triples. The Utopia project [21] proposed the assignment of semantic comments to articles.

A growing number of scientific publications, especially in the biomedical area, such as the BMJ (British Medical Journal) and the JAMA (Journal of American Medical Association), have been using structured abstracts [22] as a way to optimally extract the contents of articles.

3 Materials and methods

- The domain of biomedical sciences was chosen because scientific articles in this area follow a strict formal pattern in their texts, with sections defined according to a standard called IMRAD (Introduction, Method, Results, and Discussion).
- 89 articles in biomedical sciences were analyzed to develop the model with the aim of identifying the semantic elements of scientific methodology, reasoning patterns, and sequencing that combine these elements.

Articles analyzed comprise 3 groups.

- articles from two outstanding Brazilian research journals, 20 articles from the Memórias do Instituto Oswaldo Cruz, which has its scope mainly in Microbiology, (published during the period 1999-2004), 20 articles from the Brazilian Journal of Medical and Biological Research (published during the period 1998-2004).

- 20 articles about stem cells were also analyzed (published during the period 1994-2004). Stem cells, as an emerging research area in rapid development, were chosen expecting to find articles reporting

important discoveries. The articles analyzed were selected from three reviews which present stem cell research development in a historical perspective, pointing out the advances in research, thus of special interest for our work.

- 29 articles from the Albert Lasker Basic Medical Research Award 2006 key publications were analyzed. This last group is of special interest to the objectives of this research because the articles report, step by step, the rise of new scientific discovery, the discovery of telomerase enzyme since 1978 - the first article - to 2001 - the last article of this group. The analysis of this group of articles was guided by an article [23] by the three winners of Lasker Award 2006 which comments the steps toward the discovery of telomerase enzyme.
- Each article was analyzed in 4 steps: (1) identify patterns of reasoning developed throughout the article; (2) identify the main conclusion posited by the author in the text; (3) format the claim made in the conclusion as a relation according to the proposed knowledge representation format; and (4) tentatively map each element of the relation to concepts in the UMLS/UMLS SN. Mapping is achieved by comparing terms in the relation extracted in step 3 to MeSH/UMLS terms indexing the article in PubMed records.
- A prototype of a submission interface to an electronic journal system was developed, which formats the natural language text of conclusions of articles submitted by authors as semantic relations; this was developed using MetaMap [24], a program that processes biomedical texts to identify terms from the UMLS Thesaurus.

4 Results and discussion

We have been working for years [25] on the development of a semantic model of electronic publishing. The aim of this model is to achieve a semantically richer content surrogate of biomedical articles in a program "understandable" format. Such a knowledge representation format allows programs to extract "inferences" about the knowledge content of articles, enabling semantically powerful content retrieval and management relative to current bibliographic IR Systems. The proposed model comprises two components: a semantic content model and a Web interface for authors self-publishing and self-submitting articles to a journal system. The semantic content model *extends* conventional bibliographic record models, which comprise conventional descriptive elements such as authors, title, bibliographic source, and publication date together with content information such as keywords or descriptors. Scientific claims made by authors in their papers are represented as *relations* between two different phenomena or between a phenomenon and its characteristics [26]. Our study also includes the development of a prototype system of a Web author's submission interface to a journal system, which implements the model [27] and the use of the general framework proposed to identify discoveries in scientific papers based on two aspects: their rhetoric elements and formats and by comparing the content of the conclusion of articles with terminological data banks [28]. This last aspect corresponds to step 4 of the analysis process described in section 2 and to the task performed by authors as illustrated in Figure 5.

The following figure shows an overview of the semantic model of electronic publishing, which includes the following components: the Web interface to a system for the submission of articles to electronic publications, the Database, the public Web knowledge base, and the Discoveries identification tool.

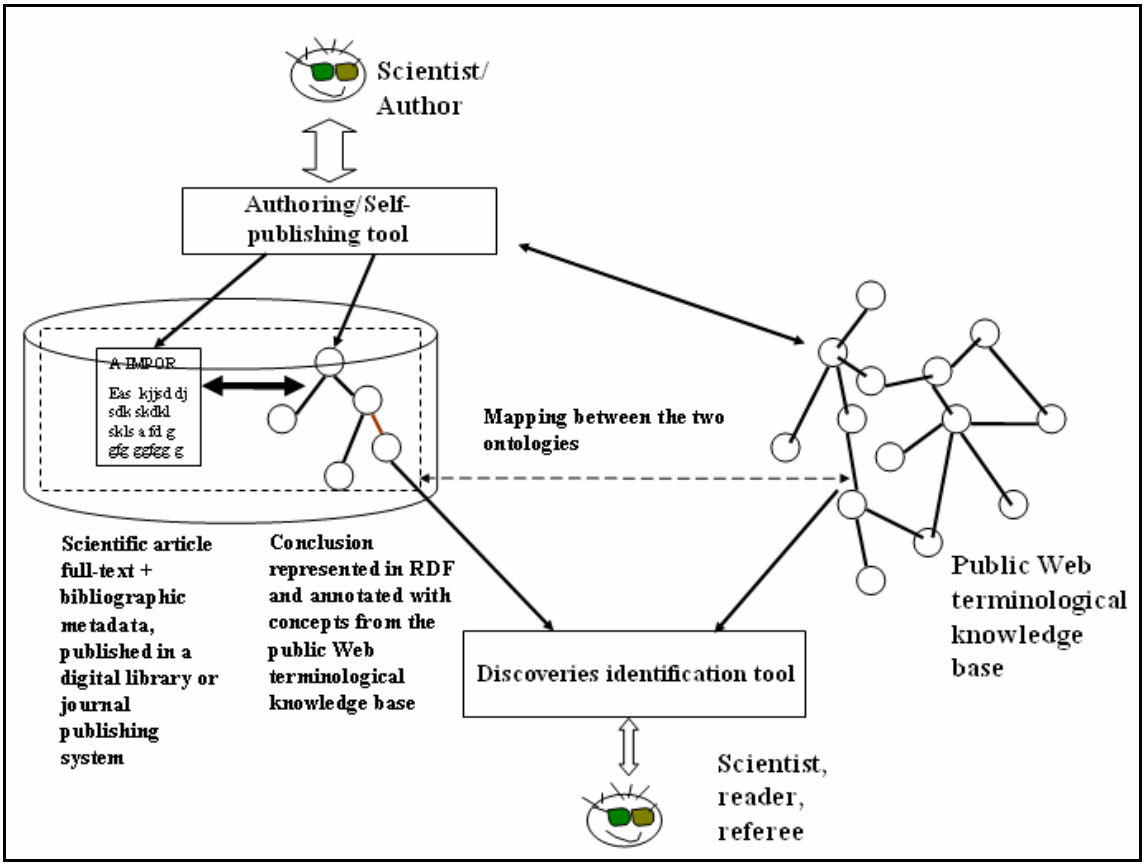

Fig. 1. Overview of the components of the semantic publication model

4.1. A semantic content model for electronic publishing

Relations are the core of the proposed knowledge representation scheme. A relation has the form of an Antecedent (a concept referring to a phenomenon), a Semantic Relation, and a Consequent (a concept referring to a phenomenon or a characteristic of the phenomenon in the Antecedent). A Semantic Relation may be a specific Type_of_relation such as "causes," "affects," or "indicates," or a (has/have) characteristic relation. Examples of knowledge representation according to this schema are the following:
- Tetrahymena extracts (Antecedent) have (Characteristic) a specific telomere terminal transferase activity (Consequent).
- Telomere shortening (Antecedent) causes (Type_of_relation) cellular senescence (Consequent).

Relations may also appear in different semantic elements throughout the article text, such as in the Problem that the article addresses; in a *Question*, in which either one of the two *relata* or the type of relation is unknown; in the *Hypothesis;* or in the *Conclusion*. Frequently, the Conclusion also poses new Questions.

Questions, *Hypothesis*, and *Conclusion* are the semantic elements comprising the proposed model. They are the elements related to the knowledge content of an article, which we aim to identify and record in a machine-processable format. The *Conclusion* is an essential semantic element that synthesizes the knowledge content of an article. In the scope of a recently published article, it is provisional knowledge; however, it is at least guaranteed by the experiment reported in the article. Semantic elements such as *Questions* and *Hypothesis* are important because they enable the evolution of a claim to be determined. Other elements have rhetoric functions, as extensively discussed in [29] and [30], or serve to describe methodological options, the experiment performed, its context, or the obtained results more clearly.

In Biomedical Sciences, there are some standardized methodological procedures, such as PCR (polymerase chain reaction), and some standardized contexts where experiments can take place, for example, in humans (e.g., children, women, embryo), rats, etc.

The semantic elements that comprise the proposed record model are as follows:

- the problem the article is addressing and the **question** derived from it,
- an **antecedent**,

- a **type_of_relation** (holding the semantic of the relation in a domain, for example, in Biomedical Sciences),
- and the **consequent**.

The **antecedent** and **consequent** may be two different phenomena or a phenomenon and its characteristics.

A possible empirically controlled **experiment** with the aim of observing the phenomenon described and specifics of experimental articles are divided into

- **results** – tables, figures, and numeric data reporting the observations made;
- **measure** used;
- a specific **context** where the empirical observations take place, subdivided into:
  - **environment** – a hospital, a daycare center, a high school,
  - a geographical **place** where the empirical observations take place,
  - **time** when the empirical observations occur,
  - a specific **population** – pregnant women, early born babies, mice – in which the phenomenon occurs,
  - **conclusion** – a set of propositions made by the author as a result of his/her findings.

A **conclusion** corroborates totally or partially the **hypothesis** of an article or negates it. A **conclusion** may also be conclusive or not yet conclusive.

In every analyzed article, concepts found in the antecedent, type_of_relation, and consequent were tentatively mapped (and will be annotated in the future web authoring/publishing tool) to concepts taken from the UMLS. Not all elements are present in all articles.

Articles differ in the way they are built around previously stated hypotheses—those stated by authors other than the author of the current article, or new, original hypotheses, i.e., those stated by the author of the current article. Articles may also differ by the existence of a documented experiment or simply theoretical considerations comparing previously stated hypotheses. We found four patterns of reasoning in the analyzed articles: *theoretical articles*, which employ abductive reasoning and *experimental articles,* which may simply be *exploratory* or employ *inductive* or *deductive* reasoning.

Theoretical-abductive (TA) articles analyze different, previous hypotheses, showing their faults and limitations and proposing a new hypothesis; the reasoning is as follows:

*A **problem** is identified, with the following aspects and data…;*

*The **previous hypotheses** (from other authors) are not satisfactory to solve the problem due to the following criticism…;*

*Therefore, we propose this **new hypothesis** (original), which we consider a new pathway to solve the problem.*

Experimental-inductive (EI) articles propose a hypothesis and develop experiments to test and validate it; the reasoning is as follows:

*A **problem** is identified, with the following aspects and data…;*

*A possible solution to this **problem** can be based on the following new **hypothesis…**;*

*We developed an **experiment** to test this **hypothesis** and obtained the following **results**.*

In experimental-inductive articles, a **conclusion** may be mainly one of these alternatives: it corroborates the hypothesis, refutes it, or partially corroborates the hypothesis. However, in some cases, the Conclusion is not one of the former; it simply reports intermediate, and not conclusive, results toward the hypothesis corroboration.

Experimental-deductive (ED) articles use a hypothesis proposed by other researchers cited by the articles' author and apply it to a slightly different context; the reasoning is as follows:

*A **problem** is identified, with the following aspects and data…;*

*In the literature, the **previous hypotheses** (by other authors) have been proposed…;*

*We choose the following **previous hypothesis…**;*

*We enlarge and recontextualize this **hypothesis**; we develop an **experiment** to test it in this new context…;*

*The **experiment** shows the following **results** in this new **context**.*

Experimental-exploratory (EE) articles are not usually hypothesis driven; their objective is to acquire knowledge about a poorly understood scientific phenomenon by performing an **experiment**; the reasoning is as follows:

There is a phenomenon that is poorly understood in a scientific domain.

*We developed an **experiment** that permits the identification of the following characteristics of this phenomenon.*

Within the group of 89 articles that were analyzed, we classified 27 as experimental-inductives (EI), 44 as experimental-deductives (ED), 15 as experimental-exploratories (EE), and 3 as theoretical-abductives (TA).

These basic semantic elements of scientific articles are interrelated and structured. Together with the corresponding bibliographic metadata and article full-text, they form richer article surrogates in machine-understandable formats and constitute single digital objects stored in a digital library or electronic journal publishing system.

The different reasoning semantic elements and reasoning procedures discussed previously can be formalized in the Model of Knowledge in Articles (MKA), as illustrated in Figure 2 with the hierarchy of classes and properties.

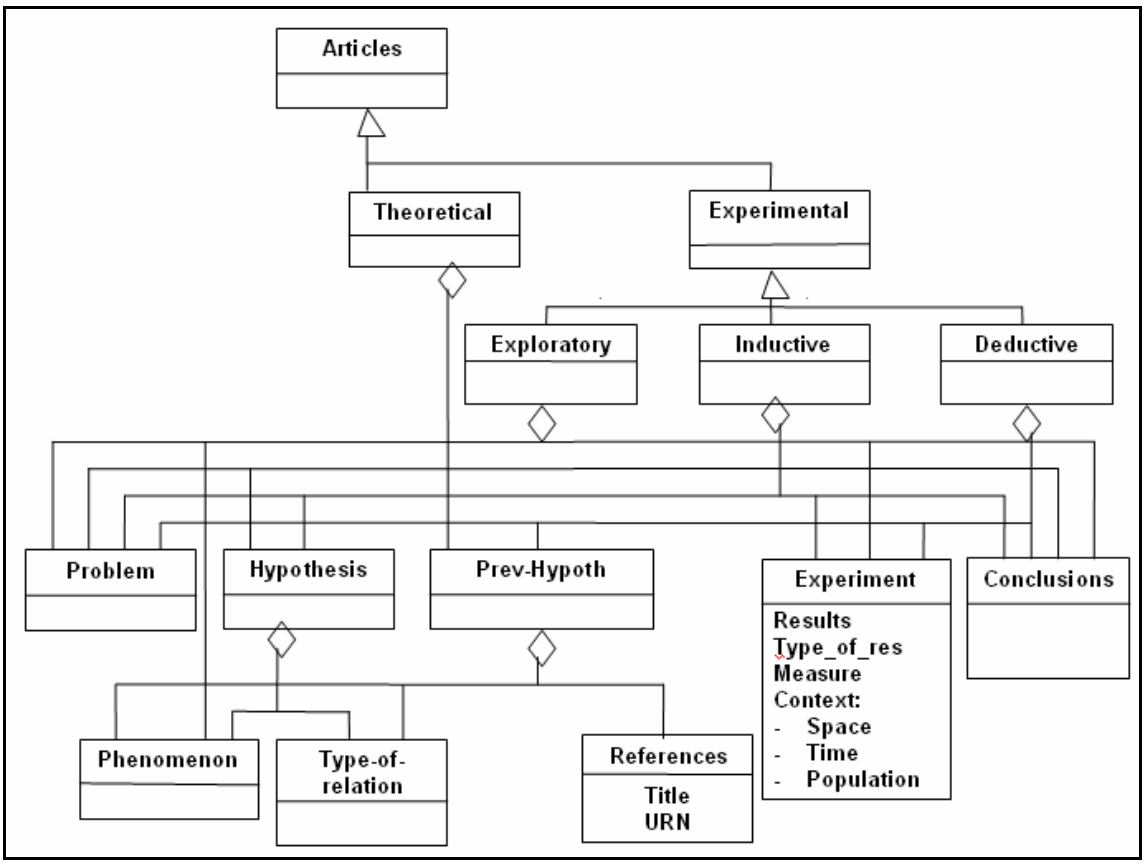

Fig. 2. MKA: model of knowledge representation in articles

The proposed knowledge representation framework enables the following types of queries to a semantic information retrieval system:

- Which other articles have hypotheses suggesting HPV as the cause of cervical neoplasias in women?

- Which articles have hypotheses suggesting other causes of cervical neoplasias different from HPV in women?

- Which articles have hypotheses suggesting HPV as the cause of cervical neoplasias in groups different from women?

- Which articles have hypotheses suggesting HPV as the cause of pathologies different from neoplasias?

- Which articles have hypotheses suggesting HPV as the cause of cervical neoplasias in different contexts (not in women from the Federal District, Brazil)?

The model also enables queries that may indicate new discoveries, for example, new causes for cellular senescence:

- Which experimental-inductive articles propose (Antecedent?) causes (Type_of_relation) for cellular senescence (Consequent) that are not mapped to UMLS concepts?

- Is there any confirmation of the hypothesis that “Several aspects of both the structural and dynamic properties of telomeres (Antecedent) led to the proposal that telomere replication

involves (Type_of_relation) nontemplate addition of telomeric repeats onto the ends of chromosomes (Consequent)?" [31]?

- Who and when first maintained that "the RNA component of telomerase (Antecedent) may be directly involved in (Type_of_relation) recognizing the unique three-dimensional structure of the G-rich telomeric oligonucleotide primers (Consequent)" [32]?

Previous examples show how the proposed knowledge representation schema may improve semantic retrieval and the use of knowledge in different and unpredicted contexts.

The implementation of the model described in a Web submission interface to an electronic journal system poses the following different challenges: representing the model, even partially, in a machine-understandable format, and extracting and formatting a relation from the article conclusion. We address these challenges as follows. We opt for an initial and partial implementation of the model of content in articles in RDF as it enables semantic retrieval using SPARQL. The following figure shows as the conclusion "telomere replication (Antecedent) involves (Type_of_relation) a terminal transferase-like activity which adds the host cell telomeric sequence repeats onto recognizable telomeric ends (Consequent)," found in [32], which is implemented in RDF format.

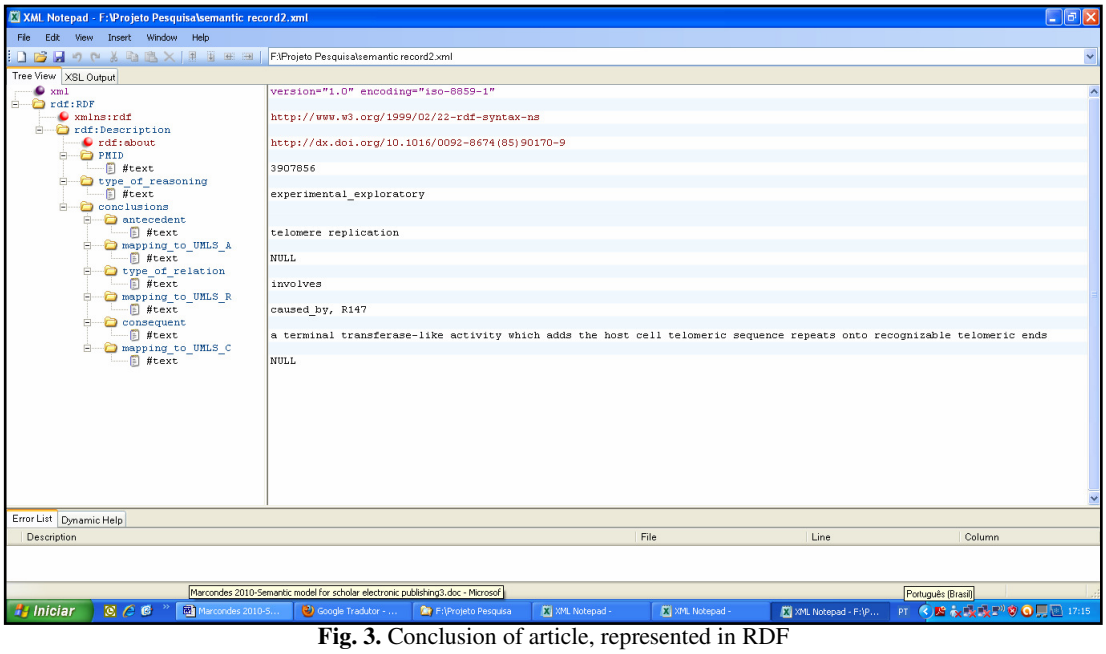

Fig. 3. Conclusion of article, represented in RDF

4.2. Web submission interface to an electronic journal system

We developed a prototype of the submission system to evaluate the dialog with authors and the extraction routine. In the future, we plan to integrate this prototype with the PKP Open Journal System [33], an electronic journal system largely used in Brazil. In its present implementation, among the semantic elements that comprise the content model, the prototype processes only the conclusion.

This prototype processes selected parts of the text, namely, the title, abstract, keywords, introduction, methods, and results; the introduction and abstract are used to extract the objective of the article through the identification of phrases such as *objectives of our work…* and *The goal of the present work…* The author is asked by the system to enter the conclusion of the article being submitted.

The extraction routine uses a formula, which is based on the frequency of occurrence of a term in the title, abstract, keywords, method, results, and objective, to weigh terms in the conclusion in order to format it from a textual format to a relation. The syntactic components found in the conclusion with higher weights are candidates for the Antecedent and Consequent of the relation. The Antecedent and Consequent must not be consecutive. The identification of a Relation requires the use of a dictionary that relates the 54 UMLS relations to a set of verbs with the same meaning, obtained from Wordnet (2010) [34].

The systems interacts with authors as follows: (1) authors are asked to enter conventional bibliographic metadata; (2) authors are asked to upload a file with article full-text; (3) authors are asked to choose the type of reasoning used in the article, either theoretical or experimental; (4) authors are asked to validate

the article objective extracted by the system; (5) authors are asked to specify the conclusion of the article; (6) after identifying its elements, the article conclusion is formatted as a relation and authors are asked to validate the Antecedent, Relation, and Consequent prompted by the system; (7) authors are asked to map concepts in the article's conclusion to UMLS terms.

After the author validates the Relation, the system records it as an instance of the MKA according to the format illustrated in Fig. 3, together with the conventional bibliographic metadata and the article full-text.

Some of the steps described above when processing the conclusion "*The results presented herein emphasize the importance to accomplish systematic serological screening during pregnancy in order to prevent the occurrence of elevated number of infants with congenital toxoplasmosis*" are shown in the following Figures.

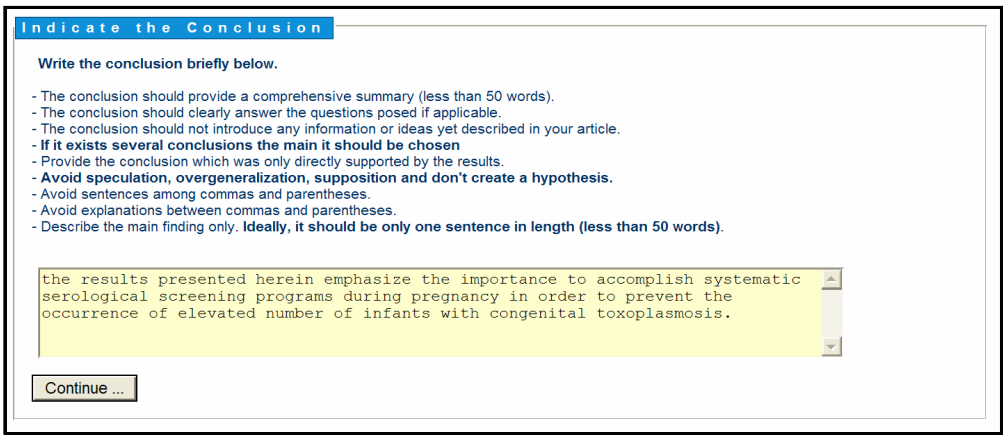

Fig. 4. Author specifies the article conclusion

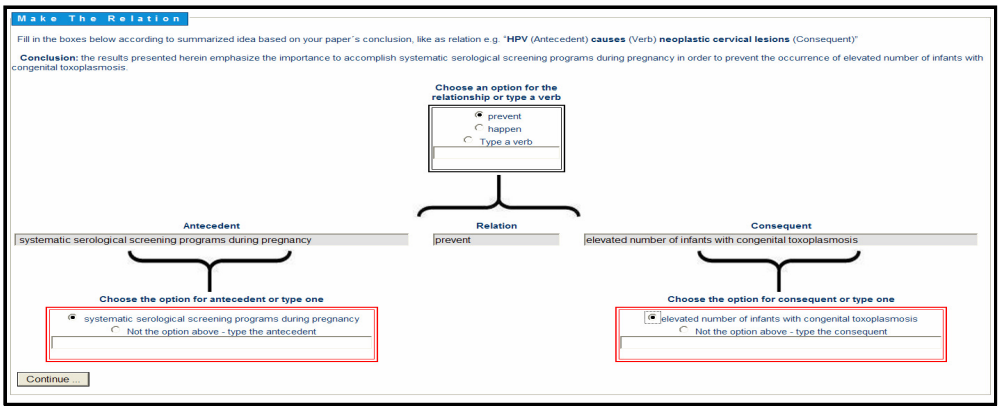

Fig. 5. The article conclusion is formatted as a relation

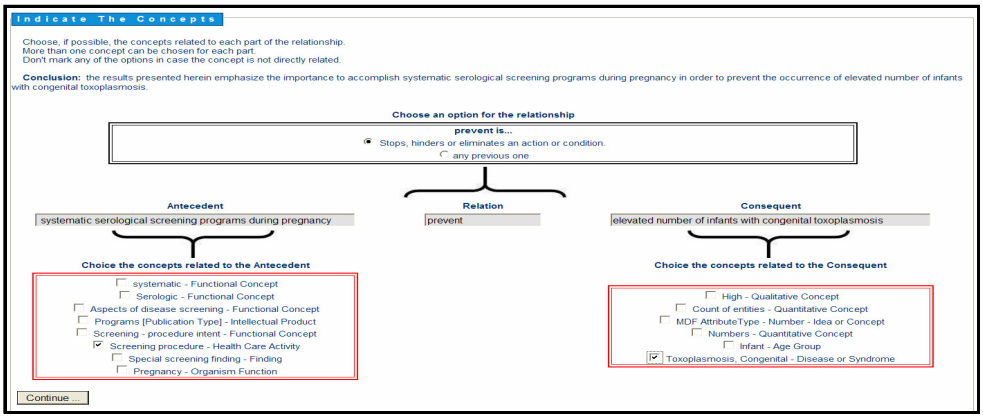

Fig. 6. Authors are asked to map concepts in the article's conclusion to UMLS terms

The prototype of the interface is in its initial phase of development. In addition to the 10 interviews, the prototype was tested with 5 of the 10 authors and in all cases, it was able to format a second relationship from the conclusion of the article.

5. Conclusions

Nowadays, researchers are accustomed to publishing and describing their papers themselves when submitting them to a digital library, conference management system, digital repository, or journal system. We consider the submission of an article to a journal system to be a privileged process during which authors are particularly motivated to clarify and disambiguate questions about their articles. The pathway that seems more feasible to reach this objective is to provide authors with an interactive interface that enables them to validate the automatic natural language processing carried out by the system. Some elements of the proposed model can be directly obtained by asking questions of the authors, such as whether the article is theoretical or experimental, whether the conclusion confirms or denies the hypotheses, and whether the article is based on the hypothesis of other authors or is original.

After the claims made by an author from anywhere in the article text, for example, the conclusion, are extracted, they will be represented in a structured form as relations. All these semantic elements can be added to conventional bibliographic elements such as the title, author, abstract, publication data, abstract, and key words, forming richer article surrogates. This knowledge content will then be represented in a standard machine-understandable format such as RDF. Articles published according to the model proposed can be interlinked and have their content annotated with an increasing number of Web public ontologies, forming a rich knowledge network. This will enable software agents to help scientists to identify and validate new discoveries in Science by comparing the knowledge content of articles with the knowledge content held in public knowledge bases such as the UMLS.

Although relations play a key role in scientific knowledge, conventional indexing languages do not take them into consideration. The inclusion of relations in knowledge representation makes an expressive difference [35] by enhancing meaning and making more precise the role of subject headings used to represent the document content.

The inclusion of articles conclusions formatted as relations to enhance article metadata is just a proposal. The prototype developed aims at testing it feasibility. The complete article record lay-out is under development.

The body of scientific literature published on the Web is becoming increasingly vast and complex. It will be necessary for scientists to have enhanced software tools in order to make inferences based on this content. Library and Information Science can go beyond conventional indexing techniques to provide fast access to full-text scientific articles. This would help scientists to directly process the knowledge content of scientific articles and to recover the reasoning that leads to a scientific discovery. The proposed model also recommends the standardization of an SkML (Scientific Knowledge Markup Language) encompassing the knowledge content of scientific articles published on the Web, as also proposed by other studies [36], [37], [38]. This opens a new perspective in scientific electronic publishing, knowledge acquisition, storage, processing, and sharing. The proposed model depends on the development of software tools that are not available yet. Our research group has not been able to fully develop the model

to the potentialities outlined here. The proposed model should, however, serve as a starting point that can be discussed and built upon by the scientific community.

Acknowledgments
This research was supported, at different times, by CNPq, CAPES, FAPERJ, and PROPPi/UFF. We would also like to thank Marília Alvarenga Rocha Mendonça, Luciana Reis Malheiros and Leonardo Cruz da Costa.

References

1. UMLS Semantic Network, http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html
2. Berners-Lee, T., Hendler, J., Lassila, O. The semantic web, Scientific American. (2001)
3. RDF Resource Description Framework, http://www.w3.org/RDF/ (accessed 10 Jan. 2007)
4. RDF Schema Specification, http://www.w3.org/TR/2000/CR-rdf-schema-20000327/
5. Gardin, J-C. Vers un remodelage des publications savantes: ses rapports avec sciences de l'information. In: Filtrage et Résumé Automatique de l'Information sur les Reseaux - Actes du 3ème Colloque du Chapitre Français de l'ISKO. Paris, Université de Nanterre-Paris X (2001)
6. OWL Ontology Web Language Overview, http://www.w3.org/TR/owl-features/
7. Renear, A. H., Palmer, C. L. Strategic reading, ontologies and the future of scientific publishing. Science 325, pp. 828--832 (2009)
8. Frohmann, B. Documentation redux: Prolegomenon to (another) philosophy of information. Library Trends 52, (3) pp. 387--407 (2004).
9. Cronin, B. Scholarly communication and epistemic cultures. Journal New Review of Academic Librarianship 9, (1) pp. 1--24 (2003)
10. Bezerman, C. Shaping written knowledge: Rhetoric of the human sciences. Madison, The University of Wisconsin Press (1988)
11. Gross, A. G. The Rhetoric of Science. Cambridge, Massachusetts; London: Harvard University Press (1990)
12. Hutchins, J. On the structure of scientific texts. In: Proceedings of the 5th. UEA Papers in Linguistics, Norwich pp. 18--39. Norwich, University of East Anglia (1977)
13. Franklin, L. R. Exploratory Experiments. In: Philosophy of Science Assoc. 19th Biennial Meeting - PSA2004: Contributed Papers, Austin, TX; 2004. Austin, Texas (2004)
14. Weinstein, J. N. 'Omic' and hypothesis-driven research in the molecular pharmacology of cancer. Current Opinion in Pharmacology 2, (4) pp. 61--65 (2002)
15. Shotton, D., Portwin, K., Klyne, G., Miles, A. Adventures in semantic publishing: Exemplar semantic enhancements of a research article. PLoS Comput. Biol. 5, (4) (2009)
16. Racunas, S. A., Shah, N. H., Albert I., Fedoroff, N. V. HyBrow: a prototype system for computer-aided hypothesis evaluation. Bioinformatics 20, (1) pp. 257--264 (2004)
17. Hunter, L., Baumgartner, W. A., Lu, Z., Johnson, H. L., Caporaso, J. G., Paquette, J., Lindemann, A., White, E. K., Medvedeva, O., Cohen, K. B. Concept recognition for extracting protein interaction relations from biomedical text. Genome Biol. 9 (Suppl 2), (2008)
18. Dinakarpadian, D., Lee, Y., Vishwanath, K., Lingambhotla, R. MachineProse: An ontological framework for scientific assertions. Journal of the American Medical Informatics Association 13, (2) pp. 220--232 (2006)
19. De Waard, A., Buckingham Shum, S., Carusi, A., Park, J., Samwald, M., Sandor, Á. Hypotheses, evidence and relationships: The HypER approach for representing scientific knowledge claims. In: Proceedings 8th International Semantic Web Conference, Workshop on Semantic Web Applications in Scientific Discourse. Lecture Notes in Computer Science. Springer Verlag Berlin, Washington DC (2009)
20. Groth, P., Gibson, A., Velterop, J. The anatomy of a nanopublication. Information Services & Use 30, pp.51--56 (2010)
21. Attwood, T. K., Kell, D. B., Mcdermott, P., Marsh, J., Pettifer, S. R., Thorne, D. Calling international rescue: knowledge lost in literature and data landslide! Biochemical Journal, Dec (2009)
22. Guimarães, C. A. Structured abstracts: Narrative review. Acta Cirúrgica Brasileira, 21, (4) (2006)
23. Blackburn, E. H, Greider, C. W., Szostak, J. Telomeres and telomerase: the path from maize, Tetrahymena and yeast to human cancer and aging. Nature 12 (10), pp.1133--1138 (2006)
24. MetaMap, http://mmtx.nlm.nih.gov/
25. Marcondes, C. H. From scientific communication to public knowledge: the scientific article Web published as a knowledge base. In: ICCC ElPub - International Conference on Electronic Publishing, Leuven, Bélgica, 2005, 9, Leuven, Bélgiun pp. 119--127. Peeters Publishing, Leuven (2005)
26. Dahlberg, I. Conceptual structures and systematization. International Forum on Information and Documentation 20, (3) pp. 9--24 (1995)
27. Costa, L. C. Um proposta de processo de submissão de artigos científicos à publicações eletrônicas semânticas em Ciências Biomédicas. Tese (doutorado), Programa de Pós-graduação em Ciência da Informação UFF-IBICT. Niterói (2010)
28. Marcondes, C. H., Malheiros, L. R. Identifying traces scientific discoveries by comparing the content of articles in biomedical sciences with web ontologies. In: 12 ISSI - International Conference on Informetrics and Scientometrics, 2009, Rio de Janeiro, v. 1. pp. 173--177. São Paulo, BIREME/PAHO/WHO, UFRJ (2009)

29. Skelton, J. Analysis of the structure of original research papers: an aid to writing original papers for publication. British Journal of General Practice, 44, pp. 455--459 (1994)
30. Nwogu, K. N. The Medical Research Paper: Structure and Functions. English for Specific Purposes 16, (2) pp. 119--138 (1997)
31. Shampay, J., Szostak, J. W., Blackburn, E. H. DNA sequences of telomeres maintained in yeast. Nature 310, pp. 154-157 (1984)
32. Greider, C. W., Blackburn, E. H. The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. Cell 51, pp. 887--898, (1987)
33. PKP Open Journal System, http://pkp.sfu.ca/
34. WordNet. A lexical database for English, http://wordnet.princeton.edu/
35. Kajikawa, Y, Abe, K., Noda, S. Filling the gap between researchers studying different materials and different methods: a proposal for structured keywords. Journal of Information Science 32, pp. 511--524 (2006)
36. Murray-Rust, P., Rzepa, H. S. Chemical Markup, XML and the World Wide Web. I: Basic principles, Journal of Chemical Information and Computer Science 39, pp. 928--942 (1999)
37. Hucka, M., Finney, A., Suro, H., Bolouri, H. System Biology Markup Language (SBML) Level 1: Structures and facilities for basic model definitions. (2003)
38. Murray-Rust, P., Rzepa, H.S. STMML. A markup language for scientific, technical and medical publishing, Data Science Journal 1, (2), pp. 128--193 (2002)

Towards New Scholarly Communication: A Case Study of the 4A Framework

Pavel Smrz and Jaroslav Dytrych

Brno University of Technology
Faculty of Information Technology
Bozetechova 2, 612 66 Brno, Czech Republic
E-mail: {smrz,idytrych}@fit.vutbr.cz

Abstract. This paper discusses the use of semantic web technology to realize the vision of future scientific publishing and scholarly communication. It introduces a novel knowledge system that builds on the popularity of social tagging and collaboration systems as well as on the idea of "Annotations Anywhere, Annotations Anytime" (hence 4A). Key technical characteristics of the realized components are presented. User experience observations and the results of a preliminary experiment are also reported.

1 Introduction

Recent years brought various proposals for changes in the model of scientific publishing and scholar communication. Most radical ones call for a complete shift from pre-publication peer review and impact factor measures towards social popularity of papers [13, 1]. Moderate views, e. g., [15, 16], consider how the success of community-specific services (arXiv[1], ACP[2]) could be replicated in other fields.

The 4A framework introduced in this paper adds a piece to the mosaic of tools that enable such changes. It can be seen as a modularization and unit-testing system in the parallel between scientific knowledge artifacts and software artifacts [6]. Pieces of knowledge on a very low granularity level (e. g., particular ideas) being annotated play the role of modules or units in this parallel. Testing corresponds to validation of annotations in collaborative knowledge building and to community appreciation of automatically generated summaries, inference-based re-uses, and other applications of the semantically enhanced resources.

Transformations in scholar communication are often motivated by benefits they could bring to a research community. On the other hand, business models applicable in the new publishing era need serious consideration [21]. Publishers obviously try to find their position in the evolving environment. It can be demonstrated by their effort to engage authors and get their feedback or to support sharing of scientific citations among researchers in the form of reference manager

[1] http://arxiv.org/
[2] http://www.atmospheric-chemistry-and-physics.net/

services (e. g., Connotea[3] or CiteULike[4]). The 4A framework is also intended to encourage the publishers to see opportunities in making money by adding value to open access data, rather than restricting the access to outcomes of research [9]. Publishers could build on their experience in organizing bodies of experts (e. g., editorial boards and review teams) able to identify the quality of the content. By means of advanced knowledge processing tools, they could extend these competences to trace expertise in specific domains, to aid the knowledge structuring process, its sharing and reuse, to support the community work around research topics, and to measure the impact of individual as well as group activities.

The framework emphasizes the role of annotation in the process of knowledge creation. It is crucial to realize that tags can be associated not only with a scientific publication, but also with any other material relevant for the research. If the experimental data referred to in a paper is made available (as required in some disciplines today) and also annotated, it is easier to verify its interpretation presented in the text. The annotation of data that is employed in experiments by other researchers also simplifies tracing its reuse and thus provides an alternative to the current state characterized by reusing and not appraising [7].

A special attention has been also paid to the means of tracking down particular ideas described in papers to the source code of computer programs described in the text. One can employ the concept of program documentation generation (for example, by means of the popular Doxygen tool[5]) and interconnect particular pieces of code with the description of their functionality on the higher level.

Last but not least, the concept of annotation is general enough to cover not only the textual material related to the presentation of a particular research result (slides from talks, blogs, tweets or other formats of messages referring to the results) but also related audio- or video recordings. For example, the proof-of-concept case study involved a material from the Speaker and Language Recognition Workshop Oddysey 2010[6] and it employed the search-in-speech service Superlectures[7]. Many ideas from referring texts are not easy to link to the material available in textual form so that working with a multimedia content is often necessary.

To summarize the directions of motivation mentioned in this section, the new scientific publishing and scholarly communication model should be accompanied by an annotation framework supporting the whole life-cycle of scientific papers – from ideas, hypotheses, identification of related research, and data collection, through setting and running experiments, implementing solutions and interpreting results, to submitting, reviewing, reflecting reviews, preparing final versions, publishing complementary material, getting feedback, discussing content and reflecting previous results in a new work. Annotations on all levels pave

[3] `http://www.connotea.org/`
[4] `http://www.citeulike.org/`
[5] `http://www.stack.nl/~dimitri/doxygen/`
[6] `http://www.speakerodyssey.com/`
[7] `http://www.superlectures.com/odyssey/`

the way for shared knowledge understanding. The social dimension of ubiquitous annotations of knowledge artifacts can also bring immediate benefits to research communities in terms of better models of fine-grained impact characteristics. The elaborated annotation system helps to pinpoint an expertise in particular fields. Researchers can also benefit from instant gratification – the annotation is immediately available for others, it can be shared and re-used in other contexts.

This paper discusses the use of semantic technologies to realize the above-mentioned vision. The name of the annotation framework – 4A – refers to the concept of "annotation anywhere, annotation anytime". Key features of the realized solution are introduced in Section 2. A proof-of-concept implementation of the 4A annotation client and its application in experiments are presented in Section 3. The paper concludes by discussing the directions of future work.

2 The 4A framework

Semantic search and other advanced functionality of the future web offer clear benefits for the semantic annotation of knowledge artifacts. On the other hand, the semantic enrichment of resources presents a tedious work for users. It is therefore crucial to lower the barrier to annotate by means of immersing the annotation into everyday work of users.

We address this requirement by defining a general interaction schema and protocols that can be supported in various contexts. An ultimate goal of the 4A framework is therefore to let users annotate naturally in any application used. The current implementation focuses mainly on the textual resources and implements the functionality of server components as well as several clients.

From the annotation support perspective, it is crucial to distinguish two types of environments users work in viewer- and editor applications. As viewers do not modify the source content, changes in annotations need to be transferred only. On the other hand, it is tricky to synchronize editing annotation sessions as the changes in text may invalidate annotations.

To serve both types of clients, a general annotation exchange protocol has been defined. The 4A synchronization server implements the protocol and enables coordinated work of 4A clients. An annotation extension to general Javascript-based editors has been developed as the first 4A client. A PDF reader add-on and a Firefox browser extension are being implemented.

The 4A framework goes beyond the current practice of simple keyword tagging and knowledge structuring curated in advance. It introduces an intuitive knowledge structuring schema of structured tags [5]. In the experiments, it has been successfully used for describing necessary conditions, conflicting views, and comparison patterns. The 4A clients present structured tags in the form of a relation attribute tree.

Unfortunately, none of the annotation formats applied in existing tools is general enough to suit our purposes without modifications. Even Annotea [10] – a format resulting from a W3C initiative to standardize annotations – cannot

express relations among structured annotations. That is why we could not simply reuse an existing format. The 4A format extends Annotea by introducing structured annotations with attributes of various types, embedded annotations and interrelations among annotations. It is based on RDF where the subject is always the annotation. It includes: annotation ID (URI), type, time of creation, its author, URI of an annotated document (or its server copy), XPath to an annotated textual fragment, its offset, length and textual content, annotation content and a specification of annotation attributes. The RDF Schema corresponding to the RDF model is available at `http://nlp.fit.vutbr.cz/annotations/rdfs/annotation-ns.rdf`.

The position of an annotated fragment is given by the path in the document object model (DOM), the offset and the size. The representation is robust to changes in general formatting. It is usually not necessary to process the whole document. For example, a web page boilerplate and other parts that are not in a DOM node on the path to an annotated fragment will be ignored.

The annotation types form a hierarchical structure. They include common comments (a note, a description) and basic types of entities (a thing, a person, etc.). Users can add new types and create complex type hierarchies.

The URI of an annotated document identifies a copy of the document that is stored on the server. The annotation process starts with a synchronization step in which the client sends the document URI and its content to the server. The server returns the URI of the local copy of the document which will be used in annotations. This procedure enables annotating documents that the server could not access directly. Processing the original document on the server side also enables removing irrelevant attributes (e. g., session ID) and applying changes to the correct version of a document as the stored version is updated together with all annotations at every access.

A new annotation interchange protocol has been defined for the communication between 4A clients and the server(s). In addition to actual annotations, it can be used for simple authentication, synchronization of annotated documents, subscription to annotations from defined sources, annotation suggestions, interchange of knowledge structures (annotation and attribute types) and various annotation-related settings.

The protocol enables two-way asynchronous communication between clients and servers. If a user adds an annotation, the server sends it immediately to all other users that annotate the same document and are subscribed to a given channel (defined by an author, a group or an annotation type). Changes of annotation types, of the document content and of relevant settings are distributed immediately as well.

Messages are defined in XML. They can be therefore easily sent over various protocols on a lower level and parsed on the client side. It is also possible to combine the messages into one XML and make the communication even more efficient.

Session management messages include the protocol version negotiation, logins and log-outs. Subscription management enables specifying what types of

annotations from what sources should be sent to a particular client. Annotations can come from another user or a URI representing an automatic annotation server, user group or other general source (e. g., an external service).

The server gets a copy of the current version of an annotated document by means of a document synchronization process. If it is the first time the document is sent, the server just stores it. If there is already a copy of the document, the server compares the new version with the stored one and updates the stored version together with all annotations. If the new version impacts no previous annotation, the operation is confirmed instantly. If there are annotations that could be invalid, the server informs users who can consequently correct possible errors.

To support clients that are not able to work with structured texts, the server can linearize the text from documents. The client transforms the document to plain text and sends it to the server. If there is a structured form of the document at the server, it is linearized and compared to the received version. If they are the same, it is possible to start the annotation process. The server will then adapt all incoming annotations for the structured version of the document. The linearization also enables cooperating with clients working with other structured formats than the server. The client will adapt positions to the linearized text, the server will adapt it to its structured form. This way it is possible to annotate the same text synchronously e. g. in HTML and in PDF.

3 Proof-of-Concept Experiments

As a proof of concept, we employed the current version of the JavaScript annotation editor in experiments . The tool is a universal component which can be easily integrated into various JavaScript-based editors such as TinyMCE[8]. It implements the functionality of a 4A client – it enables editing complex annotations, synchronizes tagging with the server side and presents annotation suggestions provided by information extraction components.

The client makes the annotation process manageable. The type of annotation can be specified in a text field. Instant search in type names is supported – the input serves as an intelligent filter. It is also possible to browse the hierarchical structure of types. This reduces diversity of annotation types.

It is possible to add new attributes to identified relations. A name, a type and a value are associated with each attribute. The selection of an attribute type is similar to that of the annotation type. The way information is presented also corresponds to the types. An attribute can be of a simple data type, of an extended type (e. g., a geographic location), or of an annotation type. Simple and extended data types are added as new branches of the type tree. If an annotation type is utilized, it is possible to choose one of the existing annotations (an annotation reference) or to create directly a nested annotation.

It is possible to select more textual fragments in a document. Annotating more fragments simultaneously enables identifying all occurrences of a relation

[8] `http://tinymce.moxiecode.com/`

in a document. When storing it, a separate annotation is generated for each fragment. The same procedure can be applied for attributes (nested annotations) with the same name, type and content which are then displayed as a list. Clients employ suggestions to facilitate the tagging process (see Figure 1 for an example).

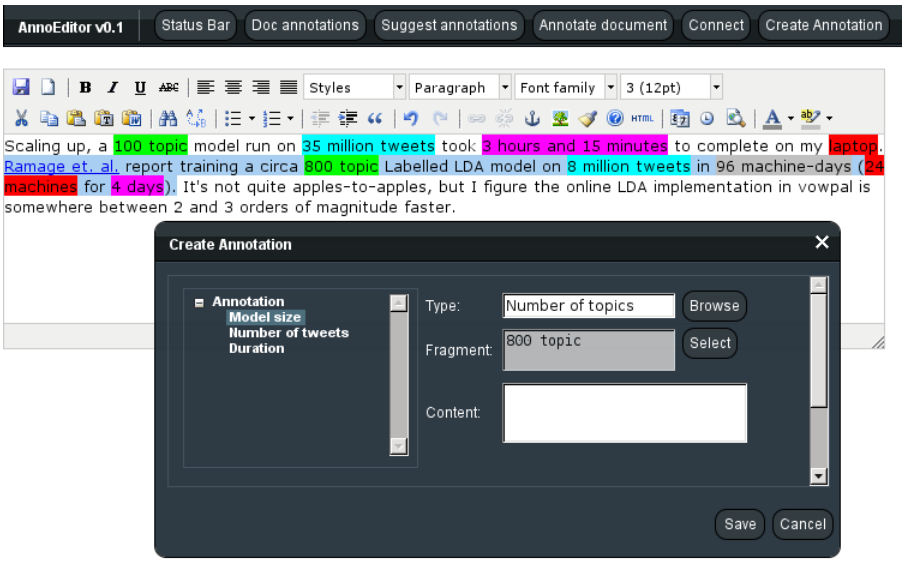

Fig. 1. Suggesting relation attributes based on the previous context

One senior researcher (the first author of this paper) and six PhD students (including the second author) participated in the annotation experiments. Source texts included not only scientific papers, but also blog messages and tweets related to specific research topics and the material referred from these sources (e. g., a particular part of a dataset, a PowerPoint presentation, a source-code file). The papers dealt mainly with natural language processing, information extraction and machine learning (correspondingly to the expertise of the participants). However, there were also general topics discussed (especially in the blog messages).

The experiments aimed at identifying particular pieces of text (selecting textual fragments) that correspond to specific types of content patterns (e. g., statements showing equivalence of two approaches or suggesting a resource for a specific group of readers). A part of the task also involved tagging the fragments supplying evidence that a particular method is really applied for a particular task (not just referred to as an alternative the paper does not really deal with). For citations, the task was to find the part of a referred text illustrating the referring context.

The primary objective of the analysis was to learn what form of tags (simple or structured ones) users prefer in given situations and how tag suggestions help to annotate the content (even if they are far from being perfect). Annotation suggestions based on automatic information extraction were switched off to not influence the answer to the first part of the question. This setting simulates specialized tagging where the set of available annotated examples is too limited to be used for learning-based methods or too complex to be described by a handful set of rules. The participants have been instructed that the annotation should be usable for future knowledge acquisition from annotated resources. Suggestions were turned on for the second part. Figure 2 shows a typical setting of the environment.

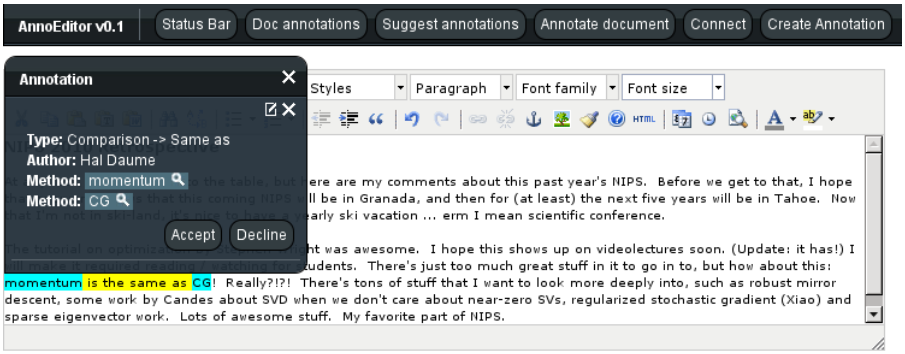

Fig. 2. Structure tag suggestions

Results of the first part (annotation form preference) are conclusive. All annotators resort to structured tags in the task of method comparison. On the other hand, they prefer flat tags over structured ones when dealing with simple annotations of individual methods ("conjugate gradient" is a "ML method"). As a side effect of the advanced functionality of the tool, it was also observed that the annotation form will probably stay unchanged if a user defines the structure of a tag for a specific task and others are able just to follow and re-use the definition.

The second part of the question turned to be hard to answer. Automatic annotation suggestions accelerate the tagging process but they also distract annotator's attention. The level of acceptance is subjective and significantly varies with respect to the precision and recall of the information extraction process generating suggestions. It is also questionable what the granularity of structured-tag suggestions should be, i. e., what information (that does not need to be correct) should be combined and potentially confirmed by one click. For example, one of the annotators pointed out that "momentum" should not be referred to as

"a method" in the context shown in Figure 2 so that he could not accept the suggestion as a whole.

4 Related Work

A lot of previous work has been done in fields related to the presented research. The general topic of future research, scientific publishing and scholarly communication is discussed in various contexts – linking scholarly literature with research data [3], advocating open access [8], demonstrating a role of social media in scholarly communication [4], etc.

First tools supporting general annotations on the web date back to mid-nineties [11, 18]. Among current solutions, Annozilla[9] is conceptually the closest to 4A browser add-ons – it is realized as an extension to the Firefox browser, tags are stored on local or remote servers, XPointer is used to identify the annotated part of a document. The annotation protocol developed in the Annotea project [10] is employed. In contrast to the 4A framework, Annozilla is intended for simple tagging only. A pre-defined set of annotation types is limited to general categories such as a comment, a question, an agreement/a disagreement. It is therefore not possible to use the tool for advanced knowledge structuring.

PREP Editor [14] and Bundle Editor [24] can illustrate innovative ideas in the area of interweaved text authoring and tagging. The former represents one of the first real-time collaborative text writing tools. Implemented annotations are limited to this functionality. The latter enables annotation structuring by means of grouping them into bundles. Annotations can be filtered, sorted etc. Zheng [23] states that the structured annotations proved to be more efficient and user-friendly than the simple ones.

Popular web-based editors and other office applications such as Google Docs[10] or Microsoft Office Live[11] show current trends in the development of collaborative tools. Even though there is either no or very limited annotation functionality, they need to be considered as "opinion makers" in terms of user interface simplicity.

Google Docs is also noteworthy as being a successor to Google Wave[12] – an envisioned distributed platform of Email 2.0 where clients should be connected to their particular servers and the servers should communicate by means of the Google Wave Protocol[13]. Compatible solutions, such as Novell Vibe[14], only started to appear when Google turned away from a strong support of the platform. Nevertheless, the vision of a distributed platform transferring user actions with a very low granularity stimulated the development of the 4A framework.

[9] http://annozilla.mozdev.org/
[10] https://docs.google.com
[11] http://www.officelive.com/
[12] https://wave.google.com
[13] http://www.waveprotocol.org/
[14] https://vibe.novell.com/

The area of social tagging is also relevant for general topics discussed in this paper. Various tools exist for the specific subdomain of collaborative scientific citation managers such as CiteULike, Connotea, Bibsonomy[15], or Mendeley[16]. The systems offer advanced support for the particular task (automatic extraction of relevant metadata, export to various formats, etc.) but they are usually not able to link references to particular pieces of text and do not explicitly deal with knowledge structuring.

On the other side of the knowledge processing support scale, there is a family of ontology, topic maps and other knowledge representation format editors. Protégé[17] defines a kind of standard, other tools such as Neon Toolkit[18], Ontolingua[19], or TopBraid Composer[20] stress the collaborative-, user support-, or integration aspects, respectively. Anchoring created knowledge structures in real data is often limited to examples or glosses that can be stored together with concepts and relations. In this context, resources resulting from semantic annotation efforts in computation linguistics, such as FrameNet[21] or OntoNotes[22], as well as enriching folksonomies by formal knowledge sources [2] are relevant for our research.

The presented work also extends the concept of semantic wikis. Several systems have been created around the idea of semantic web technologies enhancing the wiki way of content creation [12]. The 4A framework directly draws on our experience from the development of the KiWi system [17]. For example, the knowledge emergence approach based on structured tags gains from [5] resulting from the KiWi project.

Last but not least, one of the key components of the 4A framework – the information extraction module – relates to previous research on automatic knowledge extraction. In particular, we take advantage of the KiWi extraction elements [20, 19] that employ general purpose solutions such as Gate[23] or GeoNames[24]. In spite of the fact that information extraction does not form a main topic of this paper, relevant solutions for semi-automatic learning of ontologies from text such as TextToOnto[25] and OntoGen[26] or ontology-based information extraction [22] need to be mentioned as a source of inspiration.

[15] http://www.bibsonomy.org/
[16] http://www.mendeley.com/
[17] http://protege.stanford.edu/
[18] http://neon-toolkit.org
[19] http://www.ksl.stanford.edu/software/ontolingua/
[20] http://www.topquadrant.com/products/TB_Composer.html
[21] http://framenet.icsi.berkeley.edu/
[22] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T04
[23] http://gate.ac.uk/
[24] http://www.geonames.org/
[25] http://texttoonto.sourceforge.net/
[26] http://ontogen.ijs.si/

5 Conclusions and Future Directions

The 4A framework presented in this paper incorporates annotation into knowledge acquisition and knowledge sharing processes. The proof-of-concept implementation and the use of an annotation client in tagging experiments proved validity of the idea but also revealed imperfections of the model and flaws in the user interface of the realized tools. The future work will focus on removing these deficiencies. The concept of knowledge emergence and knowledge structuring needs to be refined. We will flesh out the support for social knowledge-creation processes. New information extraction modules will employ advanced machine learning methods to provide better suggestions in situations when only a few training examples are available.

The number of 4A clients will also grow. The development of the JavaScript annotation component and its integration into various web-based editors will continue. The Firefox browser extension and the PDF reader add-on will be finished. We will initiate the work on other extensions for LibreOffice and the Semantic Desktop. The annotation format will be proposed as an extension of the official W3C Annotea system.

New experiments will explore advanced tagging collaboration patterns. The question on the acceptable error-rate for the automatic suggestions will be also tackled. A higher number of participants and a more advanced setting of experiments are necessary to prove the efficiency of the knowledge structuring processes in the 4A framework. Last but not least, consistency and adequacy of the knowledge representation resulting from the use of the 4A tools will be studied.

Acknowledgements

The research leading to these results has received funding from the European Community's 7th Framework Programme FP7/2007-2013 under grant agreement number 270001 – Decipher.

References

1. Adler, B., de Alfaro, L., and Pye, I. Redesigning scientific reputation. *The Scientist 24*, 9 (Sept. 2010), 30. Online: `http://www.the-scientist.com/article/display/57645/`.
2. Angeletou, S., Sabou, M., and Motta, E. Improving folksonomies using formal knowledge: A case study on search. In *ASWC* (2009), pp. 276–290.
3. Attwood, T. K., Kell, D. B., McDermott, P., Marsh, J., Pettifer, S. R., and Thorne, D. Calling international rescue: knowledge lost in literature and data landslide! *Biochemical Journal 424*, 3 (2009), 317–333.
4. Bilder, G. Social media and scholarly communication. In *ISMTE European Conference, International Society of Managing and Technical Editors* (2010).

5. Bry, F., and Kotowski, J. A social vision of knowledge representation and reasoning. In *SOFSEM 2010: Theory and Practice of Computer Science*, J. van Leeuwen, A. Muscholl, D. Peleg, J. Pokorny, and B. Rumpe, Eds., vol. 5901 of *Lecture Notes in Computer Science*. Springer, 2010, pp. 235–246.
6. Casati, F., Giunchiglia, F., and Marchese, M. Liquid publications: Scientific publications meet the web, 2007. Online: `https://dev.liquidpub.org/svn/liquidpub/papers/deliverables/LiquidPub%20paper-latest.pdf`.
7. Enriquez, V., Judson, S. W., Weber, N. M., Allard, S., Cook, R. B., Piwowar, H. A., Sandusky, R. J., Vision, T. J., and Wilson, B. Data citation in the wild, 2010. DOI: 10.1038/npre.2010.5452.1, Online: `http://precedings.nature.com/documents/5452/version/1`.
8. Furnival, A. C. Open access to scholarly communications: advantages, policy and advocacy, 2011. Online: `http://eprints.nottingham.ac.uk/1419/`.
9. Hall, M. Efficiency and effectiveness: Digital futures in innovation, Oct. 2010. Presentation at the JISC Future of Research Conference on 19th October 2010.
10. Kahan, J., and Koivunen, M.-R. Annotea: An open RDF infrastructure for shared web annotations. In *Proceedings of the 10th International Conference on World Wide Web* (2001), ACM, pp. 623–632.
11. LaLiberte, D., and Braverman, A. A protocol for scalable group and public annotations. *Comput. Netw. ISDN Syst. 27* (Apr. 1995), 911–918.
12. Lange, C., Schaffert, S., Skaf-Molli, H., and Völkel, M., Eds. *4th Semantic Wiki Workshop (SemWiki 2009) at the 6th European Semantic Web Conference (ESWC 2009), Hersonissos, Greece, June 1st, 2009. Proceedings* (2009), vol. 464 of *CEUR Workshop Proceedings*, CEUR-WS.org.
13. Moody, G. Abundance obsoletes peer review, so drop it. `http://opendotdotdot.blogspot.com/2010/06/abundance-obsoletes-peer-review-so-drop.html`.
14. Neuwirth, C. M., Kaufer, D. S., Chandhok, R., and Morris, J. H. Computer support for distributed collaborative writing: defining parameters of interaction. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work* (New York, NY, USA, 1994), CSCW '94, ACM, pp. 145–152.
15. Nielsen, M. The future of science, 2008. Online: `http://michaelnielsen.org/blog/the-future-of-science-2/`.
16. Poeschl, U. Interactive open access publishing and public peer review: The effectiveness of transparency and self-regulation in scientific quality assurance. *IFLA Journal 36*, I (2010), 40–46. DOI: 10.1177/0340035209359573, Online: `http://www.atmospheric-chemistry-and-physics.net/pr_acp_poschl_ifla_journal_2010_interactive_open_access_publishing.pdf`.
17. Schaffert, S., Eder, J., Grünwald, S., Kurz, T., Radulescu, M., Sint, R., and Stroka, S. KiWi – A platform for semantic social software. In *Proceedings of the 4th Workshop on Semantic Wikis, European Semantic Web Conference* (2009).
18. Schickler, M. A., Mazer, M. S., and Brooks, C. Pan-browser support for annotations and other meta-information on the world wide web. In *Proceedings of the fifth international World Wide Web conference on Computer networks and ISDN systems* (Amsterdam, The Netherlands, The Netherlands, 1996), Elsevier Science Publishers B. V., pp. 1063–1074.
19. Schmidt, M., and Smrz, P. Annotation component for a semantic wiki. In *Proceedings of the 5th Workshop on Semantic Wikis, European Semantic Web Conference* (2010).
20. Smrz, P., and Schmidt, M. Information extraction in semantic wikis. In *Proceedings of the 4th Workshop on Semantic Wikis, European Semantic Web Conference* (2009).

21. Waltham, M. Why does one size not fit all in journal publishing?, 2010. Presentation at the June 2010 Society for Scholarly Publishing meeting, slides available online: http://www.marywaltham.com/SSP_Seminar_2010.pdf.
22. Wimalasuriya, D. C., and Dou, D. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science 36*, 3 (2010), 306–323.
23. Zheng, Q. Structured annotations to support collaborative writing workflow. Master's thesis, The Faculty of Graduate Studies (Computer Science), The University of British Columbia, Dec. 2005. (accessed online March 14, 2011).
24. Zheng, Q., Booth, K., and McGrenere, J. Co-authoring with structured annotations. Department of Computer Science, University of British Columbia, 2006.