RDF Authoring in Wikis

Florian Schmedding, Christoph Hanke, and Thomas Hornung

Institute of Computer Science, Albert-Ludwigs University Freiburg, Germany
{schmeddi, hankec, hornungt}@informatik.uni-freiburg.de

Abstract Although the Semantic Web vision is gaining momentum and the underlying technologies are used in many different areas, there still seems to be no agreement on how they should be used in everyday documents, such as news, blogs or wiki pages. In this paper we argue that two aspects are crucial for the enrichment of this documents with semantic annotations: full support for RDF and close integration of the annotations with the continuous text. The former is necessary because many common relationships cannot be expressed by attribute-value-pairs, the latter reduces redundancy and enables Web browsers to help readers using the contained data. To gain further insights, we implemented an RDFa-capable extension for MediaWiki and report on improvements for wiki use cases and other applications on top of the contained data.

1 Introduction

Since the vision of the Semantic Web [1] has been described, different knowledge markup and ontology definition languages, such as RDF [2] and OWL [3] have been proposed and standardized. A recent survey has shown that these languages are mostly applied to highly-structured domains with a well-understood semantics, e.g. for drug discovery [4]. In the revisited version [5] of the original vision the authors acknowledge that the Semantic Web has not reached the expected adoption yet. We believe that this is because the lion's share of the content on the Web is only available in presentation-oriented HTML documents without any semantic markup. Therefore to reach a critical user base, a clear benefit for the users of everyday documents, such as news, blogs or wiki pages, has to be established.

So far, semantic annotations were added to HTML documents in a very informal and restricted manner with respect to the semantic complexity of the information. In our opinion these approaches still suffer from two major drawbacks: lack of expressiveness and separation of text and annotations. The goal of our approach is to have full RDF expressivity while retaining the proximity of metadata and normal textual content. The latter is especially important for reusing existing external applications or enabling third parties to make use of the semantic content in a standardized way, e.g. for extracting calendar data.

Projects such as DBpedia[1] have shown that Wikipedia[2] already contains a lot of relevant structured metadata, e.g. the population of cities, and hence is

[1] http://dbpedia.org
[2] http://wikipedia.org

an ideal candidate for the adoption of Semantic Web technologies to enrich the existing content with semantic annotations. In this paper we describe an extension for MediaWiki[3], which allows to directly embed these semantic annotations while editing the wiki article. The main features are full support of RDF, including blank nodes, and the direct embedding of the resulting annotations in the generated XHTML presentation of the wiki article.

The remainder of the paper is organized as follows. In Section 2 we introduce and evaluate semantic annotation formats with respect to our requirements. In Section 3 we describe our extension to MediaWiki and present a use case about geo-political facts of countries in Section 4. In Section 5 we discuss related work, Section 6 gives an outlook on future work, and we conclude with Section 7.

2 Semantic Annotation Proposals

In line with our proposal, all semantic annotations should be embedded in the written article to avoid the administrative overhead of maintaining separate documents (HTML and e.g. RDF/XML). Additionally, there are less redundancies and update problems with a single document which is both human- and machine-readable from a single Web address.

Currently, there are three competing proposals for annotating semantics in HTML documents: Microformats[4], eRDF[5], and RDFa [6]. For obvious reasons, RDF/XML [7] is out of the question because it is not designed to contain readable text. Because eRDF only supports a subset of RDF and the informal semantics of Microformats we chose RDFa as annotation language.

Another orthogonal approach to embedding semantics in HTML documents is GRDDL [8]. It is designed to extract RDF data from any XML document via specialized XSLT transformations. Redundancy in text and RDF data is therefore omitted, but there is no connection between text and RDF data.

3 RDFa Wiki Extension

We implemented a prototype as an extension of MediaWiki to evaluate our approach. Here, the main focus is on augmenting the existing wiki syntax to enable users to embed arbitrary RDF statements into regular articles. The syntax design especially considers the following three requirements:

1. Subject and object of a statement can be any desired URI (or blank nodes),
2. Subject and predicate should be invisible to the reader and literals should be masqueradable,
3. Single statements are made within one unit, as distributed statements are vulnerable to partial deletion, which could alter the semantics.

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

[4] http://microformats.org

[5] http://research.talis.com/2005/erdf/wiki/Main/RdfInHtml

In general, the semantic statements in our wiki extension include subject, predicate and object, although the subject is not mandatory. To annotate an existing wiki text, the user has to choose the desired object in the wiki text and place the predicate and optionally a subject in front of it. The whole semantic statement is delimited by <sem>-tags. URIs for subject and object are expressed using the common MediaWiki link notations. The predicate has to be written in CURIE [6] style, e.g. `cc:license` instead of the expanded URL `http://creativecommons.org/ns#license`. Applying these rules a statement about the external page `www.mypage.de` would look like this:

```
My homepage is licensed under <sem> [http://www.mypage.de] cc:license
[http://mylicense.org/ my own license] </sem>.
```

Omitting the subject (`[http://www.mypage.de]`) would create an equivalent statement but about the current page. In both cases the only value visible to the user is the URI, or rather the label of the object.

If the object is not denoted in link notation, the object value is interpreted as a (XML-)Literal. Literals can additionally receive a dataype and a label. This can be used to provide a date in a machine readable format, which means we masquerade the machine-readable data with an alternative representation. For example the sentence *The meeting takes place on 7th of August 2008* could be annotated in the following way, where "2008-08-07" represents the machine-readable date and *7th of August 2008* is the alternative representation:

```
The meeting takes place on <sem>[http://www.futuremeeting.com] dc:date
"2008−08−07"^^xsd:date 7th of August 2008 </sem>.
```

A further feature of our extension is the possibility of using blank nodes. For this we introduce a three bracket notation to provide a name for the blank node variable. This concept is useful for adequate modeling of n-ary relationships [9], e.g. to describe the border between two countries, where also the length of the border is of interest. An exemplary statement is given here:

```
The border between <sem>[[[border]]] mond:bordering [[Spain]]</sem> and
<sem>[[[border]]] mond:bordering [[France]]</sem>
has a length of <sem>[[[border]]] mond:length 623</sem> km.
```

Additonally the type of the subject can be classified using the `inof` attribute of the <sem>-tag:

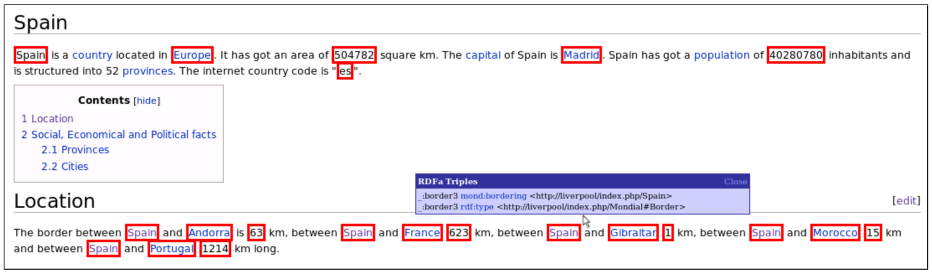
```
<sem inof="mond:Border">[[[border]]] mond:bordering [[Spain]]</sem>
```

Figure 1 shows a geographical wiki page about Spain. The JavaScript tool RDFa-Highlight[6] can be used in any browser to mark all semantically annotated areas.

4 Use Cases

Mondial [10] is a collection of political and geographical data, which covers typical concepts that we expect in a semantically enhanced version of Wikipedia.

[6] `http://www.w3.org/2001/sw/BestPractices/HTML/rdfa-bookmarklet/`



Spain

Spain is a country located in Europe. It has got an area of 504782 square km. The capital of Spain is Madrid. Spain has got a population of 40280780 inhabitants and is structured into 52 provinces. The internet country code is "es".

Contents [hide]

1 Location
2 Social, Economical and Political facts
2.1 Provinces
2.2 Cities

RDFa Triples Close

_:border3 mond:bordering <http://liverpool/index.php/Spain>
_:border3 rdf:type <http://liverpool/index.php/Mondial#Border>

Location [edit]

The border between Spain and Andorra is 63 km, between Spain and France 623 km, between Spain and Gibraltar 1 km, between Spain and Morocco 15 km and between Spain and Portugal 1214 km long.

Figure 1. Wiki page about Spain with marked semantic annotations (red). The blue box shows two statements about a blank node.

For this reason, we chose it as the basis to populate our prototype wiki with approximately 5,500 test pages.

4.1 Wiki Page Generation

We used a simplified version of the Mondial RDFS ontology[7] for the generation of the wiki pages, e.g. an excerpt for Spain is depicted in Figure 1. For the sake of illustrating the features of the ontology, we concentrated on the therein defined concepts and relationships. Obviously, we could have also enriched regular Wikipedia articles with additional semantic annotations.

Although the pages were generated according to a pre-defined template the simplified articles demonstrate the feasibility of our approach for real-world scenarios. It is also a good example of how to bootstrap semantic wikis from existing database content.

4.2 Ontology Maintenance

In contrast to other wikis, which separate the metadata from wiki articles, in our approach RDF vocabularies can be defined within the articles itself. This follows as an immediate consequence of the RDF support in our syntax. New definitions could be stated on any arbitrary wiki page or in a more structured way, using a reserved wiki category or special page. For example, the abovementioned Mondial ontology is defined in a separate article by means of our new wiki syntax. This enables the wiki community to collectively define and evolve ontologies with the same syntax used for authoring semantic articles.

4.3 Data Import

Since the RDFa standard is on its way to becoming a W3C recommendation, we expect the number of accessible XHTML+RDFa pages to constantly increase in the near future. Each of these pages could be seen as a remote information source,

[7] http://www.dbis.informatik.uni-goettingen.de/Mondial/

analogously to a SPARQL endpoint. This would enable us to use a coherent query language to both specify queries on our wiki and to include these remote sources as well in our wiki articles as dynamic data sources. An example from the Mondial theme would be to include the gross domestic product of a (future) semantically annotated version of the CIA World Factbook[8]. In this case the changes would occur only once a year, but the same general concept applies to including the most current publications of researchers in the relevant articles. Since for most data on the Web, especially homepages, it is not realistic to expect the data to be availabe via SPARQL endpoints, we expect a reasonable application area of our coherent integration approach.

5 Related Work

Similar to Semantic MediaWiki [11], we base our extension on MediaWiki. But unlike this project, our main focus was to have maximum RDF support for authors, instead of maintaining the current wiki syntax. Although this requires additional effort on the side of the authors, we believe that the benefits of the added semantics outweigh this inconvenience. For example, we support subjects different from the current page. BOWiki [12] is an extension of Semantic MediaWiki and is additionally capable of representing n-ary predicates but is restricted to a specialized biological domain. Kaukolu [13] also supports subjects different from the current page but has no full support for blank nodes. IkeWiki [14] is geared towards knowledge engineers and provides a sophisticated user interface and ontology reasoning. Our approach is geared towards shallow ontologies [5] and regular users. OntoWiki [15] offers a visual editor for easy editing of RDF content and provides semantically enhanced search strategies. Its main focus is on the acquisition of instance data and knowledge engineering projects. We are more interested in enriching normal wiki texts with embedded semantics. SweetWiki [16] also uses RDFa to embed semantics directly in the articles. However, their major focus is on providing keywords, or so-called tags, for specific articles or objects, e.g. images, inside the article. Our focus is on authoring complex RDF relations between several entities within the article. Finally, the internal structure of the Maariwa [17] wiki is based on an ontology meta model, i.e. each page either represents a class, an individual or a set of individuals. The annotations are then interpreted as properties of the class or individual, respectively. To query information they introduce a proprietary query language called MarQL. As discussed in Section 4.2 we propose a less rigid approach to specifying ontologies within arbitrary articles.

6 Future Work

To allow non expert users to formulate complex annotations in their articles it is crucial to provide an intuitive and easy-to-use editing environment. Inspired by

[8] `https://www.cia.gov/library/publications/the-world-factbook/index.html`

the successful WYSIWYG principle, we are currently working on a rich internet application for editing and annotating semantic Wiki articles in an integrated fashion.

Up to now, the available visualization tools for contained RDFa annotations in (X)HTML pages are rather limited. Given the specific Wiki content of Mondial we envision a more sophisticated graphical presentation of contained RDFa annotations. Currently, we are exploring different approaches of how to better support the user in browsing and understanding the contained annotations.

Additionally, the close proximity of semantic annotations to the textual content opens the door for new information retrieval applications, e.g. to combine keyword-based searches with semantic enhancements. By only querying the contained RDF data in a triple store, unannotated text is not considered. At the moment, we are investigating the impact of combining the approach by [18] with RDFa annotated pages.

Measurement units are currently not considered, but could be handled similar to the Semantic MediaWiki [11] proposal. An open question is how to handle articles in different languages about the same concept: should the annotations be shared between the different versions or does each language belong in a separate semantic unit? This a general question, which is not specific to our approach, and is relevant for each semantic wiki to some extent.

7 Conclusion

We have shown a semantic extension for MediaWiki and how it helps to improve the application of semantic wikis as well as the benefits of the directly embedded annotations for other applications, e.g. for developing semantic-aware search engines. Additionally, our approach could contribute to the proliferation of semantically enriched content on the Web, especially with a higher-level editing environment that hides the syntactic details of the wiki syntax. If the advantages of these semantic annotations would be visible to and demanded by end users the willingness of authors to employ these techniques would increase significantly. We believe this is possible in the near future due to the standardization of RDFa by the W3C and expect a wide adoption and support in Web browsers as well as innovative uses by other third party tools.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American (May 2001)
2. Manola, F., Miller, E.: RDF Primer. http://www.w3.org/TR/rdf-primer (2004)
3. Smith, M.K., Welty, C., McGuinness, D.L.: OWL Web Ontology Language Guide. http://www.w3.org/TR/owl-guide/ (2004)
4. Feigenbaum, L., Herman, I., Hongsermeier, T., Neumann, E., Stephens, S.: The Semantic Web in Action. Scientific American **297** (December 2007) 90–97
5. Shadbolt, N., Berners-Lee, T., Hall, W.: The Semantic Web Revisited. IEEE Intelligent Systems **21**(3) (July 2006) 96–101

6. Adida, B., Birbeck, M., McCarron, S., Pemberton, S.: RDFa in XHTML: Syntax and Processing. `http://www.w3.org/TR/rdfa-syntax/` (2008)
7. Beckett, D.: RDF/XML Syntax Specification (Revised). `http://www.w3.org/TR/rdf-syntax-grammar/` (2004)
8. Connolly, D.: Gleaning Resource Descriptions from Dialects of Languages (GRDDL). http://www.w3.org/TR/grddl/ (2007)
9. Noy, N., Rector, A.: Defining N-ary Relations on the Semantic Web. `http://www.w3.org/TR/swbp-n-aryRelations/` (2006)
10. May, W.: Information extraction and integration with Florid: The Mondial case study. Technical Report 131, Universität Freiburg, Institut für Informatik (1999)
11. Krötzsch, M., Vrandecic, D., Völkel, M., Haller, H., Studer, R.: Semantic Wikipedia. Journal of Web Semantics **5** (2007) 251–261
12. Backhaus, M., Kelso, J., Bacher, J., Herre, H., Hoehndorf, R., Loebe, F., Visagie, J.: BOWiki – A Collaborative Annotation and Ontology Curation Framework. In: CKC. (2007)
13. Kiesel, M.: Kaukolu: Hub of the Semantic Corporate Intranet. In Völkel, M., Schaffert, S., eds.: SemWiki. (2006)
14. Schaffert, S.: IkeWiki: A Semantic Wiki for Collaborative Knowledge Management. In: WETICE. (2006) 388–396
15. Auer, S., Dietzold, S., Riechert, T.: OntoWiki - A Tool for Social, Semantic Collaboration. In: ISWC. (2006) 736–749
16. Buffa, M., Gandon, F.L., Ereteo, G., Sander, P., Faron, C.: SweetWiki: A Semantic Wiki. J. Web Sem. **6**(1) (2008) 84–97
17. Landefeld, R., Sack, H.: Collaborative Web-Publishing with a Semantic Wiki. In: CSSW. (2007) 23–34
18. Bast, H., Chitea, A., Suchanek, F.M., Weber, I.: ESTER: Efficient Search on Text, Entities, and Relations. In: SIGIR. (2007) 671–678