Schema.org for the Semantic Web with MaDaME

Csaba Veres, Eivind Elseth

Department of Information Science and Media Studies,
Postbox 7802, University of Bergen, 5020 Bergen, Norway
csaba.veres@infomedia.uib.no, eivind.elseth@student.uib.no

Abstract. Schema.org is a high profile initiative to introduce structured markup into web sites. However, the markup is designed for use cases relevant to search engines, which limits their general usefulness. MaDaME is a tool to help web developers to annotate their web pages with schema.org annotations, but in addition automatically injects semantic metadata from SUMO and WordNet. It is unlike previous tools in that it assumes no knowledge of the metadata standards. Instead, users provide disambiguated natural language terms, and the tool automatically picks the most appropriate metadata terms from the different vocabularies.

Keywords: schema.org, wordnet, semantic web, markup, search

1 Introduction

Schema.org was launched on June 2, 2011, under the auspices of a powerful consortium consisting of Google, Bing, and Yahoo! (they were subsequently joined by Yandex). They established the http://schema.org web site whose main purpose is to document an extensive type schema which is meant to be used by web masters to add structured metadata to their content. In a sense schema.org provides extended semantics for rich snippets1 with the motivation that markup can be used to display more information about web sites on the search page which may result in more clicks, and perhaps higher rankings in the long run.

The schema was designed specifically for the use cases developed by the search engines, and both the semantics and preferred syntax reflect that choice. In terms of semantics, the schema has some non-traditional concepts to fulfill its role. For example there is a general class of *Product* but no general class for *Artifact*. There are also odd property ascriptions from the taxonomy structure, so, for example, Beach has *openingHours* and *faxNumber*. These oddities exist because, we are told, they reflect what people are predominantly looking for when they perform a search. In terms of syntax, there is a very strong message that developers should use the relatively new Microdata format, designed specifically for the schema, rather than the vastly more popular RDFa web standard [1]. The choice is dictated by simplicity, because Microdata has just those elements required for the schema. But this choice is unfortunate

[1] http://goo.gl/RAJy8

because it makes metadata from schema.org incompatible with many other sources of metadata like Facebook's OGP.[2] [2] lists five key reasons why RDFa Lite 1.1 should be the preferred syntax over Microdata. RDFa is feature equivalent to Microdata, and it is supported by all major search crawlers including Facebook, while Microdata is not. For the purposes of expressing schema.org, RDFa is no more complex than Microdata. But most importantly from the perspective of general semantic markup, RDFa is designed to naturally mix vocabularies while Microdata makes it much more difficult to do so. Thus if annotating web pages with multiple vocabularies is the desired goal, then RDFa Lite 1.1 is the best choice.

MaDaME (Meta Data Made Easy) is a markup tool developed for two specific purposes. First, it must to help web developers who were not familiar with the schema.org to mark up their web sites as easily as possible. This is important because the idiosyncratic nature of the schema can make concepts hard to navigate. It is especially important if a web developer wants to mark up a site for which there is no existing type in schema.org. For example a web master might be designing a web site about caves for tourists to visit, but schema.org does not have a type for *cave*. We wanted to help developers find the best markup in these cases, without requiring them to study the schema itself. The second important motivator was to make the markup episode as fruitful as possible, since it is not easy to motivate people to provide structured data about their web site. This means the markup should be useful in as many use cases as possible. We achieve this by producing RDFa markup and mixing different vocabularies to describe the same object. While there are existing efforts to provide tool support for schema.org markup, including a tool from Google,[3] all of them require some knowledge of the schema, and none of them provide rich markup for a more general semantic web.

2 MaDaME

MaDaME has at its core a mapping file between WordNet word senses and schema.org types. WordNet can for our purposes be regarded as a comprehensive electronic dictionary which defines word senses through numerous relationships to other words [3]. Web developers simply look up the word which expresses the content of their site, and they are given the best matching schema.org markup. Obviously not all words will have direct mappings to schema.org, so we also have an algorithm to infer the best match for those.

To import a page into the web app the user will write the URL of the web site he wants to mark up into the URL input field. The page will then be loaded into the web app after some preprocessing. The preprocessing consists of commenting out scripts and iframes which might not run correctly. The user then selects words, phrases, or images to tag by highlighting them on the page. When a word item has been highlighted, its possible senses in WordNet are retrieved. The user picks one of these senses by clicking on it. In fig. 1 we can see the word *ridge* highlighted, and the corresponding disambiguation options. The sense the user picks is sent back to the server for map-

[2] http://ogp.me
[3] http://goo.gl/7DGr5D

ping to schema.org, as well as a selection of other ontologies. So far we have only implemented SUMO [4] and WordNet itself. The Schema.org mappings can be further refined by filling out the properties defined by the schema, using a popup form.

When the users finish marking up the document they are given a link to a newly created webpage containing their original page plus the meta data they have created. In most cases where the web site is simple HTML there will be no need to manually modify any code. From here they can save the document and upload it to their own server.

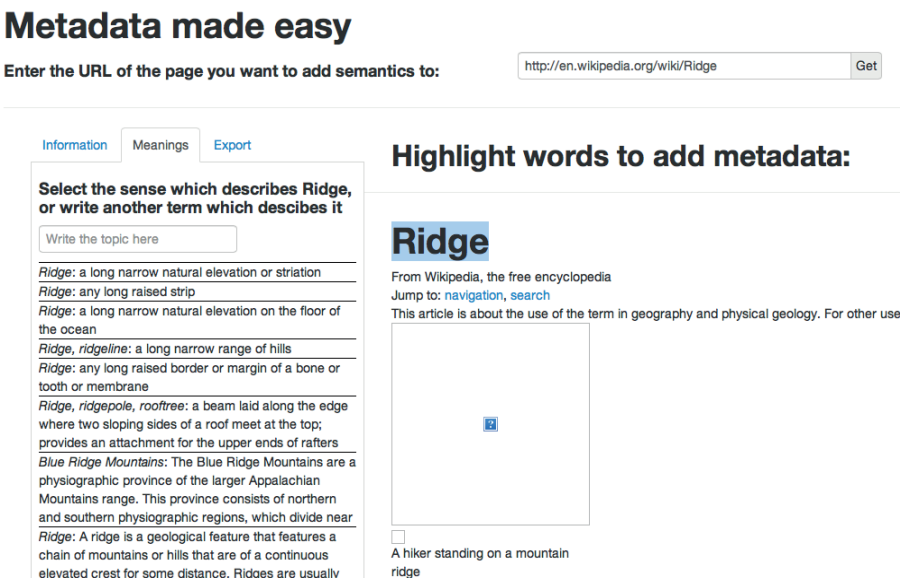

Fig. 1. A screenshot of MaDaME with options for *ridge* on the left of the screen

All of the markup is in the RDFa Lite 1.1 syntax, which is the current W3C recommendation,[4] and has the necessary features to handle multiple namespaces and multiple types elegantly.

The algorithm for finding markup for the selected senses is in two stages. The first stage is to build an extended tree of WordNet senses. This is done by using a perl library (the WordNet::QueryData library from CPAN) which is capable of querying the WordNet database. We have written a script that when given a WordNet sense will find the hypernyms of the sense (more general senses), and all the hyponyms (more specific) of these ancestor nodes. We call this the *mapping tree*, which intuitively contains all the words in the semantic neighbourhood of the original word.

In the second stage we find mappings for the user selected synsets. If a direct mapping to schema.org exists then this is simply added to the markup. For novel words we

[4] http://www.w3.org/TR/rdfa-lite/

use mappings for the closest available related sense from the *mapping tree* which does have a direct mapping. We tried several versions of the mapping algorithm, and the most successful one turned out to be a simple depth-first traversal of the *mapping tree* until a sense is found with a direct mapping to the schema. For a simple example, consider the concept *ridge* which is not represented in schema.org. The correct sense of wn:ridge has the hypernym wn:geological_formation, which has a direct mapping to schema:Landform. Therefore *ridge* is marked up as schema:Landform. SUMO has direct mappings for a very large number of WordNet senses and *ridge* has a corresponding mapping in SUMO as sumo:UplandArea, so the concept *ridge* would acquire mappings schema:Landform as well as sumo:UplandArea. More generally, any vocabulary that is mapped to WordNet could be used to provide metadata. In future releases we plan to provide facilities for advanced users to incorporate their own mapping files to an ontology of their choice.[5]

3 Results

We performed an automatic evaluation of 4350 random nouns in WordNet to see how they mapped to schema types, by measuring the average depth of the mapped type in the schema.org taxonomy. The result was a somewhat disappointing 0.689, which means that most words were mapped to schema:Thing or one of its immediate specialisations.

To test how this compares to real world usage we sampled a set of five web sites that had used schema.org markup. We ended up with a restaurant review from the Telegraph, a tour operators customer feedback page, a tourist agency home page, the home page of a marketing company and a movie review sites review of a film. When we manually added markup by selecting key words in the text we achieved 100% agreement. While this is clearly a small study, it does suggest that the schema.org markup we will see "in the wild" will represent concepts from the top nodes of the type hierarchy. The relatively shallow mappings may be a reflection of the schema itself, rather than a criticism of our mapping algorithm.

4 Related Work

There are existing approaches for annotating web pages with semantic markup, especially schema.org. These can broadly be categorised as manual or automatic annotation tools.

The schema.rdfs.org web site links to a number of publishing tools[6]. The two major form-based tools, Schema Creator and Microdata Generator, both provide a forms based interface for entering detailed properties, not unlike the MaDaME interface. However in these tools the web author must find the appropriate schema types by

[5] The tool can be tried at http://csaba.dyndns.ws:3000.
[6] http://schema.rdfs.org/tools.html

browsing a sub set of the most common types that are presented in these tools. They both differ from our approach because they expect the author to make decisions about which schema types to use. Similarly, major content management platforms like Drupal, Joomla!, WordPress and Virtuoso provide mechanisms for adding schema.org types to their content.

Amongst automatic annotation tools, [5] presents a tool that can add schema.org types automatically, but only to web pages about patents. Their approach uses underlying domain knowledge to extract key terms and a patent knowledge base to generate structured microdata markup for web pages. It remains to be seen if this approach could scale to web sites in general.

5 Conclusion

Schema.org is a promising initiative from the search engines in that it exposes structured metadata to a vast new audience of web developers. However, this requires some learning of the syntax and vocabulary of the particular markup, which could limit the breadth of metadata that will appear from web developers. MaDaME is a tool that helps web masters use the schema because it removes the requirement to learn a new vocabulary and syntax, while providing the necessary markup. The markup can be extended to other proprietary standards like Facebook's OGP, so web sites could be annotated with both standards at no extra effort. But we see MaDaME's most important contribution as one to the semantic web effort because it piggybacks on the major search engine backed initiative, to include markup from popular ontologies that can be used for diverse semantic applications.

References

1. P. Mika and T. Potter, "Metadata Statistics for a Large Web Corpus," WWW2012 Workshop on Linked Data on the Web (LDOW '12), Lyon, France, 16-Apr-2012. Online Available: http://ceur-ws.org/Vol-937/ldow2012-inv-paper-1.pdf [Accessed: 11-Jul-2012].
2. M. Sporny, "Mythical Differences: RDFa Lite vs. Microdata | The Beautiful, Tormented Machine," manu.sporny.org. Online Available: http://manu.sporny.org/2012/mythical-differences. [Accessed: 17-Jul-2013].
3. G. A. Miller, "WordNet: a lexical database for English," Communications of the ACM, vol. 38, no. 11, Nov. 1995, pp. 39–41
4. I. Niles and A. Pease, "Towards a Standard Upper Ontology," Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS '01), ACM, 2001, pp. 2–9.
5. A. Norbaitiah and D. Lukose, "Enriching Webpages with Semantic Information," Proc. Int'l Conf. on Dublin Core and Metadata Applications 2012, Sep. 2012, pp. 1–11