Semantic Markup for Web Applications

Martin Dostal, Dalibor Fiala, and Karel Ježek

Open Science Index, Computer and Information Engineering Vol:7, No:4, 2013 publications.waset.org/15156.pdf

***Abstract*—**In this paper we would like to introduce some of the best practices of using semantic markup and its significance in the success of web applications. Search engines are one of the best ways to reach potential customers and are some of the main indicators of web sites' fruitfulness. We will introduce the most important semantic vocabularies which are used by Google and Yahoo. Afterwards, we will explain the process of semantic markup implementation and its significance for search engines and other semantic markup consumers. We will describe techniques for slow conceiving RDFa markup to our web application for collecting Call for papers (CFP) announcements.

***Keywords*—**Call for papers, Google, RDFa, semantic markup, semantic web, Yahoo.

I. Introduction

In this paper we describe current methods for the addition of semantic markup [1] to the website and we will explain them via examples. We have created a web agent for collecting Call for Papers (CFP) announcements and we would like to show ways of publishing this information in a machine readable way. We will explain the significance of this semantic marking and describe the importance of this step for search engines.

Nowadays, many sites are growing and creating much more interesting content. Additional information is getting better every day and popularity is growing. But popularity, in the form of visitor numbers, is a very uncertain factor because search engines are not able to find exact and correct answers. Current search engines mostly use the popularity of the website in the form of a page score. This is a relatively good measure in long-term planning and searching. Satisfactory static information can be found this way, but this approach is not good enough for temporary information. The temporary value of information can, for example, be an event, product information, review, video, discussion, map position, news, documents, or their combination. Each of these information sources can be usable only if it is up to date. The content of the paper is as follows.

We will describe methods for semantic markup creation in section II. Section III and IV are devoted to the biggest search engines, Google and Yahoo, and to their semantic markup support. In section V we will discuss differences between the semantic markup concepts used in these search engines and we will deal with reasons for semantic markup as presented in [1] and their advantages and disadvantages.

II. Semantic Markup

In this section we discuss semantic markup methods. First of all, we should start with a definition of the word 'semantic'. Probably the best definitions are from Wordnet: 'relating to meaning', 'study of meaning'. Tim Berners-Lee describes the Semantic Web [6] as an approach to expressing information in a machine processable form.

Having interesting data in a user readable form, the simplest way is adding the semantic markup to the existing content. Therefore, this content will be accessible in a user and machine readable way. We will not discuss the methods of basic XHTML content markup; we will discuss the methods for expressing temporary information only. The basic approaches use the descriptive power of the XTHML tag set without inventing a new format like RSS; therefore, we do not need any special software to work with that. These approaches are microformats [4] and RDFa [1].

The main difference between these approaches is the way of using XHTML attributes for the storage of metadata information. The microformats use only *class* attributes, however, RDFa uses more descriptive methods for metadata expression. The microformats provide a number of vocabulary-specific syntaxes. However, RDFa provides a more generic semantic markup embedding syntax, which is vocabulary independent. RDFa uses these XHTML attributes: *about*, *resource*, *instanceof*, *property*, *content*. Attributes like *rel* and *href* can be applied to all elements, not just for links. The use of RDFa for semantic markup has been widely discussed in recent literature, e.g. [7]-[18].

III. Google's RDFa Support

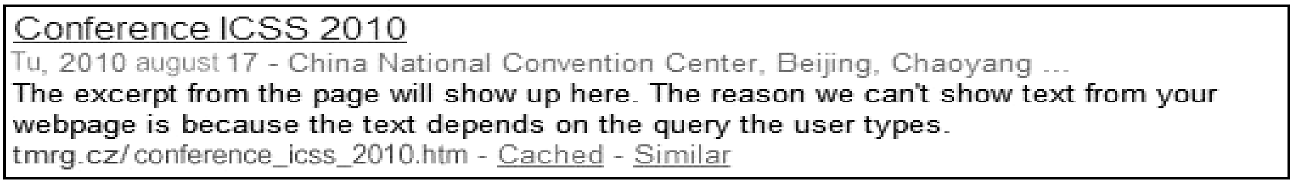
This section introduces Google's RDFa [3] support. Google uses the semantic web in a different way to others. For Google, the Semantic Web is just a source of structured information, which can be used to improve the search accuracy. The main idea of the semantic web, in the way of RDF extensibility, is missing. Google supports only a few vocabularies which are a useful source of information, while other vocabularies are totally ignored. This support is better than nothing, but Google will in the end have to support more vocabularies.

M. Dostal is with the Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 30614 Plzeň, Czech Republic (phone: +420-37763-2479; fax: +420-37763-2401; e-mail: madostal@kiv.zcu.cz).

D. Fiala is with the Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 30614 Plzeň, Czech Republic (e-mail: dalfia@kiv.zcu.cz).

K. Ježek is with the Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 30614 Plzeň, Czech Republic (e-mail: jezek_ka@kiv.zcu.cz).

The current Google RDFa support consists in using vocabularies for: *reviews*, *people*, *business and organizations*, *events*, *recipes*, and *video*.



Conference ICSS 2010
Tu, 2010 august 17 - China National Convention Center, Beijing, Chaoyang ...
The excerpt from the page will show up here. The reason we can't show text from your webpage is because the text depends on the query the user types.
tmrg.cz/conference_icss_2010.htm - Cached - Similar

Fig. 1 Google search preview – RDFa event support

The following example is given to the semantic markup of an event with information about a conference. Interesting information is marked with RDFa. The first part is a definition of the vocabulary used for the *Event*:

```
<!-- Definition of used vocabulary – Event -->
<div xmlns:v=http://rdf.data-vocabulary.org/#
typeof="v:Event">
```

Next is basic information about the event – an event name called *summary* and event description:

```
<!-- Name and description of the event -->
<span property="v:summary">ICSS 2010</span>
<span property="v:description">The 7th International
Conference on Cognitive Science (ICCS2010) will be held on
August 17-20, 2010, at the China National Convention Center
in Beijing.</span>
```

We can assign a related image to the event:

```
<!-- Image related to the event -->
<div class="image"><img
src="http://www.iccs2010.org/images/iccs2010-header.jpg"
rel="v:photo" /></div>
```

The most important information is about the beginning and end of the event:

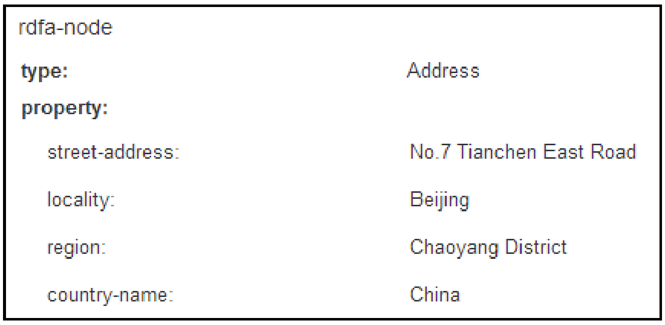
```
When:
  <span property="v:startDate"
    content="2010-08-17">August 17</span> —
  <span property="v:endDate"
    content="2010-08-20">August 20</span>
Where:
```

We can use nested entities for additional information, for example, about the location or organization:

```
<!-- Nested entities for location and organization -->
<span rel="v:location">
<span typeof="v:Organization">
<span property="v:name">China National Convention
Center</span>,

<!-- Address – street, city, country, region: -->
<span typeof="v:Address">
<span property="v:street-address">No.7 Tianchen East
Road</span>,
<span property="v:locality">Beijing</span>,
<span property="v:region">Chaoyang District</span>,
<span property="v:country-name">China</span>
```

Google search preview of this example is shown in Fig. 1. There is the *title* of the website, which should be the same as the name of the *Event*. There is the start date and location of the event on the second line. A short description of the event is generated based on the stored information in Google index. This example has not been indexed yet, so there is no event description. The corresponding RDFa node structures for address and event are shown in Fig. 2 and Fig. 3, respectively.



rdfa-node

type:	Address
property:	
street-address:	No.7 Tianchen East Road
locality:	Beijing
region:	Chaoyang District
country-name:	China

Fig. 2 Google search – Address type RDFa node structure

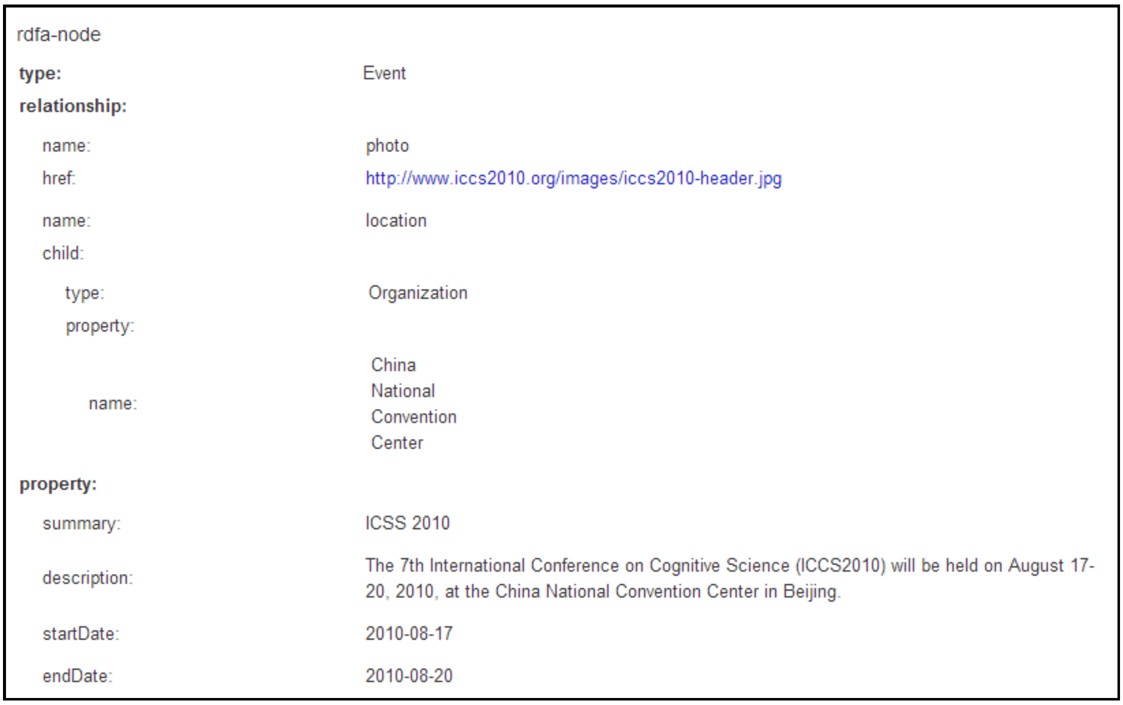
IV. Yahoo's RDFa Support

Yahoo! Search supports RDFa [5] and makes this information available to the public via SearchMonkey. SearchMonkey is Yahoo! Search's open platform based on metadata. Metadata are displayed in the form of standard enhanced results and can be used for event specific searches, for example.

The main difference between Yahoo and Google is in the openness to the most popular vocabularies. Google uses only its own vocabulary. Moreover, Yahoo is able to use some of the most popular existing vocabularies such as *Good Relations* [2]. It makes the Yahoo! Search engine more effective and open to well-known standards.

SearchMonkey supports these types of structured data:

- *Product* – information about a product including current and sale price, image, specification, structured data about the manufacturer, and reviews. The main difference between Yahoo and Google RDFa support is in the number of marked items. Google does not use these types of prices, structured data about the manufacturer is replaced by brand name, etc.



rdfa-node

type:	Event
relationship:	
name:	photo
href:	http://www.iccs2010.org/images/iccs2010-header.jpg
name:	location
child:	
type:	Organization
property:	
name:	China National Convention Center
property:	
summary:	ICSS 2010
description:	The 7th International Conference on Cognitive Science (ICCS2010) will be held on August 17-20, 2010, at the China National Convention Center in Beijing.
startDate:	2010-08-17
endDate:	2010-08-20

Fig. 3 Google search – Address type RDFa node structure

- *Video* – description, license, thumbnail image and information about the size and type of the content.
- *Discussion* – based again on the open vocabularies. For example: *foaf*, *sioc*, *dc*, *media*, *vcard,* etc. Google has no support for discussions, it has only *reviews*.
- *Local* – business, organizations and points of interests. Public vocabularies are again used: *vcard*, *comment*, *review,* etc and their vocabulary *commerce* is only for additional information. Google has a vocabulary for *businesses and organizations* withonly some basic structured fields like name, URL, address, telephone and GPS position.
- *Event* –Yahoo uses generally known vocabularies like vcard and xmlns: rdfs. Their own vocabulary, called *commerce,* is used for storing very detailed additional information, for example, about *parking options*, *opening hours*, *attire* and *type of cuisine*. Google uses only their own vocabulary and has nothing like these additional options.
- *Games* – support for Flash games only, but Google has nothing like this.
- *News* – based on public vocabularies like: *dc*, *sioc*, *vcard,* etc. Discussion is included.
- *Documents* – support for documents. We can add information about the author, license and media type.

Example of RDFa implementation of *Event* for Yahoo! Search. This implementation is for our website and contains information about a conference. We will show only the main differences:

The biggest difference is in the number of used vocabularies. Google uses only one vocabulary, their own. Yahoo uses many popular vocabularies:

```
<div typeof="vcal:Vevent"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:vcal="http://www.w3.org/2002/12/cal/icaltzd#"
  xmlns:vcard="http://www.w3.org/2006/vcard/ns#"
  xmlns:review="http://purl.org/stuff/rev#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:commerce="http://search.yahoo.com/searchmonkey/
commerce/">
```

The basic information looks very similar. This RDFa is for Yahoo – based on generally known vocabularies:

```
<span property="rdfs:label vcal:summary">ICSS
2010</span>
```

```
<span property="vcal:description rdfs:comment">The 7th
International Conference…</span>
```

This RDFa is for Google. It uses only their own vocabulary:

```
<span property="v:summary">ICSS 2010</span>
<span property="v:description">The 7th International
Conference </span>
```

Information about dates is almost the same as in RDFa for Google. RDFa for Yahoo:

```
<span property="vcal:dtstart" datatype="xsd:dateTime"
   content="2010-08-17">August 17</span> —
<span property="vcal:dtend" datatype="xsd:dateTime"
   content="2010-08-20">August 20</span>
```

V. Conclusion

First we explained methods for marking content with semantic tags. These tags can be added in the form of microformats or in the form of RDFa. The structured content can be used by search engines and can bring new visitors or potential customers to our site.

Search engines are able to mark different types of content, for example, products, events, and locations. If the content is marked in a machine readable way, search engines can classify and filter content by this type. It can bring new possibilities for users, who will be able to find any cultural event through their favourite search engine without checking many different websites. We will be able to plan our lives more effectively and update our calendars directly or automatically from these search results. We will be able to find e-shops selling interesting products and we will be able to compare reviews from different sources.

We discussed two big search engines, Google and Yahoo, and we compared their differences. Yahoo RDFa support is based on open, generally known vocabularies; on the other hand, Google prefers its own vocabulary specification. This makes Yahoo the leader as far as its way of analyzing and processing structured data is concerned.

In our future work, we would like to further concentrate on Yahoo! BOSS [19] - the successor of Yahoo! Search's SearchMonkey – and Yahoo! Content Analysis [20], both of which provide application interfaces for various services using semantic markup, such as user location detection and named entity recognition, among others.

Acknowledgment

This work was supported by the European Regional Development Fund (ERDF), project 'NTIS - New Technologies for Information Society', European Centre of Excellence, CZ.1.05/1.1.00/02.0090.

References

[1] T. Berners-Lee, W3C. URL: http://www.w3.org/DesignIssues/Semantic.html, Design Issues.

[2] Good Relations vocabulary. URL: http://www.heppnetz.de/projects/goodrelations, GoodRelations: The Web Ontology for E-Commerce.

[3] Google. URL: http://www.google.com/support/webmasters/bin/answer.py?hl=en &answer=99170, Google webmaster central - Rich snippets.

[4] R. Khare, and T. Çelik, "Microformats, A Pragmatic Path to the Semantic Web," *CommerceNet Labs Technical Report 06-0.* January 2006.

[5] Yahoo!. URL: http://developer.yahoo.com/searchmonkey/developer.html, SearchMonkey: Developer Overview.

[6] W3C – Semantic Web. URL: http://www.w3.org/RDF/FAQ, Semantic Web Frequently Asked Questions.

[7] S. Peroni, and F. Vitali, "Annotations with EARMARK for arbitrary, overlapping and out-of order markup," in *Proceedings of the 2009 ACM Symposium on Document Engineering (DocEng'09)*, Munich, Germany, pp. 171-180, 2009.

[8] J. L. Navarro-Galindo, and J. Samos, "Manual and automatic semantic annotation of Web documents: The FLERSA tool," in *Proceedings of the 12th International Conference on Information Integration and Web-Based Applications and Services (iiWAS2010)*, Paris, France, pp. 542-549, 2010.

[9] A. Kohlhase, M. Kohlhase, and C. Lange, "Dimensions of formality: A case study for MKM in software engineering," *Lecture Notes in Artifical Intelligence (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6167, pp. 355-369, 2010.

[10] M. Samwald, E. Lim, P. Masiar, L. Marenco, H. Chen, T. Morse, P. Mutalik, G. Shepherd, P. Miller, and K.-H. Cheung, "Entrez neuron RDFa: A pragmatic Semantic Web application for data integration in neuroscience research," *Studies in Health Technology and Informatics*, vol. 150, pp. 317–321, 2009.

[11] M. Pereira, and J. A. Martins, "aRDF: A plugin to expose RDFa semantic information using Grails," in *Proceedings of the 6th Euro American Conference on Telematics and Information Systems (EATIS 2012)*, art. no. 6218057, Valencia, Spain, 2012.

[12] J. L. Navarro-Galindo, and J. S. Jiménez, "Flexible range semantic annotations based on RDFa," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6121, pp. 122–126, 2012.

[13] K. Hyppönen, M. Alonen, S. Korhonen, and V. Hotti, "XHTML with RDFa as a semantic document format for CCTS modelled documents and its application for social services," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7117, pp. 229-240, 2012.

[14] T. Krishna Chaitanya, and J. Ganesh, "Role of mashups, social networking platForms and semantics in revolutionizing web integration: Key insights and enterprise implications," in *Proceedings of the 17th Americas Conference on Information Systems 2011 (AMCIS 2011)*, pp. 1917-1923, Detroit, USA, 2011.

[15] X. Bai, "Addressing the RDFa publishing bottleneck," in *Proceedings of the 20th International Conference Companion on World Wide Web (WWW 2011)*, pp. 331-335, Hyderabad, India, 2011.

[16] A. Haller, "ActiveRaUL: A model-view-controller approach for semantic web applications," in *Proceedings of the 2010 IEEE International Conference on Service-Oriented Computing and Applications (SOCA 2010)*, art. no. 5707158, Perth, Australia, 2010.

[17] R. A. Buchmann, A. Mihaila, R. Meza, "Semantic processing based on eye-tracking metrics," *WSEAS Transactions on Computers*, vol. 8, no. 10, pp. 1701-1710, 2009.

[18] F. Schmedding, M. Schwaibold, K. Simon, "Pattern-based annotation of HTML-streams," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5554, pp. 893-897, 2009.

[19] Yahoo! URL: http://developer.yahoo.com/boss/, Yahoo! Boss Search API.

[20] Yahoo! URL: http://developer.yahoo.com/contentanalysis/, Yahoo! Content Analysis API.