The Adoption of Semantic Annotations of Products in Web Shops

Goran Matošević

***Abstract*—In this paper we explore the use of semantic web technologies to mark up data in web pages in the most popular web shops today. Semantic annotations i.e. insertion of structural data within existing HTML document is done for the purpose of enabling aggregating data from various sources that may use different web applications, portals and web search engines. While there are a few studies about the global use of these technologies, there is no any existing detailed research in the field of e-commerce. Also the emergence of new standards (Microdata) and ontologies (GoodRelations and Schema.org) in the field of e-commerce requires more detailed insight into the actual use of the above mentioned. Results of this study show the current situation in this field, and can serve to web shops as a guide for the use of semantic annotation but also to scientists for further research.**

***Index Terms*—Semantic web, semantic annotations, e-commerce, product search, microdata, microformats.**

I. Introduction

Product search on the web is becoming increasingly important due to the constant increase of online shopping. Also the number of Internet stores and products that are offered in them has become larger, and for users in search of the product is more difficult to decide where and which model of product to buy [1]. This situation creates several problems. There is a variety of identical products described in different ways in different Internet stores. Furthermore, the products have been described in different languages. This leads to difficulties in the aggregation of data about the products which various portals and search engines could use. Currently the user seeking a particular product on the web using Internet search engine by entering keywords in the search field, should examine i.e. visit several websites from the search results in order to obtain the information that is requested, or choose a web shop where he wants to make a purchase considering the cost, shipping terms, and user comments. Search is done using keywords and search results are pages that contain those keywords. There is no possibility of parameter's type of searching i.e. "The blue Nike sneakers that cost between $ 50 and $ 100 (USD)." This type of search uses only a few specialized Internet stores such as *Shopping.com* and *Shopzilla.com*, but these search engines only support basic product properties [2]. In order for search engines like *Google*, *Yahoo* and *Bing* could return results of parameter's inquiries over all indexed Internet stores, they need a system of ontology and semantic annotation of products in HTML. Also, if we want to build a portal that aggregates data from multiple Internet stores, it is necessary to find a solution that will easily be able to withdraw products, compare them and aggregate. Solution is the semantic web and tagging data (semantic annotation) within HTML using a predefined vocabulary.

Manuscript received September 30, 2013; revised January 22, 2014.

G. Matošević is with the Faculty of Economics and Tourism, University of Pula, Croatia (e-mail: gmatosev@unipu.hr).

II. Semantic Web and Annotation of Content

Semantic web is a new generation of web that adds meaning to data on the web. Today's web is understandable to humans, but not understandable to machines. Semantic web extends the current one by making informations meaningfully for humans and computers, enabling the description of contents in machine-readable form [3], [4]. Semantic web makes it possible to objects from the real world to describe themselves on web pages so that they are understandable also to the machinery. These objects can be people, companies, products, etc. It is also possible to define relationships between objects. If two or more web pages in the same way denote these objects, then the integration of data is very simple too. Different technologies and pre-defined vocabularies (ontologies) can be used for semantic annotation of content. Specific domains can require specific ontologies. In this way, the semantic search engines can combine data that represent the same object from the real world from many different sources and display aggregated results.

Semantic annotation implies insertion i.e. adding of some tags and attributes in the existing HTML code. Currently there are three ways i.e. standards by which we can do this: RDFa, microformats and microdata [5]. Microdata and microformats are very easy to implement, but unlike RDFa, they provide fewer opportunities. Microdata are the latest standard recommended by the largest search engines *Google*, *Yahoo* and *Bing*. They developed (2011) unique vocabulary to indicate the most common objects such as people, products, events, reviews, etc. These are general ontology that is available on *schema.org*. Microformat has its own ontology and the method of inserting the HTML code, while RDFa recommendation of the World Wide Web Consortium (W3C) is slightly more complicated to implement.

A. Microformats

Microformats is approach of labeling i.e. inserting structural data into the existing XHTML code using the existing HTML attribute "class" that has been used to define *Cascading style sheets* (CSS) classes [6]. Microformats specification includes several predefined elements that describe different objects, such as "hCalendar" for events, "hCard" for persons or organizations, "XFN" for the relations

between people, etc.[1] In addition to adopted specification there are so called “draft specification” which consists of elements that are currently undergoing a process of standardization and should be used with caution since their syntax still can change. Among them are the elements of interest for e-commerce: “hProduct”, “hReview” and “hReview-aggregate”. They serve to describe the products and their review and rating. The following example shows how microformats are used to annotate a product with review:

```
<div class=“hproduct“>
  <span class=“fn“>Apple iPad 2</span>
  <span class=“description“>Things come alive on the stunning 9.7-inch widescreen LED Multi-Touch display of the Apple iPad 2. With WiFi support, this 16GB Apple iPad ensures you stay connected to your world all the time.
  </span>
  Sale price: <span class=“price“>$300.00</span>
  Average rating:
  <span class=“hreview-aggregate">
  <span class=“rating“> <span class="average">4.9</span>,
  based on <span class="count">99</span></span> reviews
  </span>
</div>
```

B. Microdata

Microdata are another way in which we can add semantic markups in XHTML. It is a standard that emerged from *Schema.org* initiative that in 2011 launched the largest Internet search engines *Google*, *Yahoo*, *Bing* and *Yandex* in order to show in their search results besides the usual titles and descriptions of website some additional structural information (called "rich snipets"), and thereby improve the user experience [7]. Syntax of inserting microdata involves the use of three main attributes: *itemscope*, *itemtype* and *itemprop* that can be loaded into any existing HTML tag.

To define the structural data is used *Schema.org* vocabulary that contains several classes organized in a hierarchy[2] and in the code are referred to by the full URL. Microdata unlike Microformats include wider and more advanced vocabulary and allow the use of other vocabularies by simple URL referencing. Existing class *Schema.org* can be easily expanded, and their hierarchical structure allows for a detailed description of the data. For e-commerce *Schema.org* has class "Product" and can be used in conjunction with the "Offer", "Brand", “Review", "AggregateRating" etc. The following example shows the same product from previous example labeled with Microdata and *Schema.org* vocabulary:

```
<div itemscope itemtype=“http://schema.org/Product“>
  <span itemprop=“name“>Apple iPad 2</span>
  <span itemprop=“description“> Things come alive on the stunning 9.7-inch widescreen LED Multi-Touch display of the Apple iPad 2. With WiFi support, this 16GB Apple iPad ensures you stay connected to your world all the time.
  </span>
  <span itemprop="offers" itemscope itemtype="http://schema.org/Offer">
    Sale price: $<span itemprop=“price“>300.00</span>
    <meta itemprop=“ priceCurrency“ content=“USD“ />
  </span>
  <span itemprop="aggregateRating" itemscope itemtype="http://schema.org/AggregateRating">
    Average rating <span itemprop="ratingValue">4.9</span>,
    based on <span itemprop="reviewCount">99</span> reviews
  </span>
</div>
```

C. RDFa

Resource Description Framework in attributes (RDFa) is a format that allows the insertion of RDF triples in the HTML document [8]. Syntax involves the use of URL to reference the vocabulary and elements of triplets. Facebook's Open Graph Protocol is the biggest supporter of RDFa. By using RDFa we can use any vocabulary. Our third example shows the labeling of products with GoodRelations ontology in RDFa:

```
<div
    xmlns:gr="http://purl.org/goodrelations/v1#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:v="http://rdf.data-vocabulary.org/#"
    typeof="gr:Offering" about="#product_data">
    <span property="gr:name">Apple iPad 2</span>
    <span property="gr:description">Things come alive on the stunning 9.7-inch widescreen LED Multi-Touch display of the Apple iPad 2. With WiFi support, this 16GB Apple iPad ensures you stay connected to your world all the time.
</span>
    <span rel="gr:hasPriceSpecification">
    <span typeof="gr:UnitPriceSpecification">
    Sale price: $<span property="gr:hasCurrencyValue" datatype="xsd:float">300.00</span>
    <span property="gr:hasCurrency" content="USD" datatype="xsd:string"></span>
    </span>
    </span>
  <span rel="v:hasReview">
    <span typeof="v:Review-aggregate" about="#review_data">
    Average rating: <span property="v:average" datatype="xsd:string">4.9</span>,
    based on <span property="v:count" datatype="xsd:string">99</span> reviews
    </span>
  </span>
</div>
```

D. E-Commerce Ontologies

Ontologies are used for modeling knowledge from a specific domain. They allow you to add semantic meaning of the existing concepts and to define mutual relations [4]. In the area of products and services were developed general ontologies Schema.org/Product, eClassOWL and unspscOWL. Hepp in his work [9] exposes problem of lack of detailed ontology for describing products and services, and

[1] http://microformats.org/wiki/Main_Page#Specifications
[2] http://www.schema.org/docs/full.html

suggests his ontology that he calls "GoodRelations" which today is regarded as the most complete ontology for e-commerce. Schema.org/Product vocabulary already contains some elements from GoodRelations Ontology and their recommendation is that GoodRelations should be used in the annotation of products.

III. Related Work

Reference [5] analyzed and compared the approaches to semantic tagging and publishing of machine-readable data on the web. They divide the approaches as "inline" and "parallel". Inline approach refers to the insertion of extra tags into existing HTML files in order to add semantic meaning to them and thus also allow to machines easier reading of data. The authors cite three approaches i.e. technologies by which this is possible: RDFa, Microdata and Microformats. A parallel approach implies the existence of parallel files that contain only data. In this approach belongs *Linked data* as the only approach that now allows publishing of large sets of data on the Web using RDF standards.

Reference [10] investigated the use of inline semantic tagging by processing publicly available data set of downloaded websites from the projects *Common Crawl*[3] and *Web Data Commons*. [4] *Common Crawl* is a non-profit organization that collects data i.e. websites using computer programs called *Spiders*, in a manner similar to that used today by modern Internet search engines. The collected data are published on its web site, and are publicly available for everyone to download. Until today there are two such sets of data, one from 2009/2010 and the second from February 2012, which together contain 4.5 billion of web pages. *Web commons* project uses mentioned sets of data in order to pull out from them structural data, and then publish them in the RDF format. Reference [10] showed and compared the most common formats that are used in two sets of data to show change over time and trends of the use of technology for inline semantic labeling. This study included all three technologies, RDFa, Microformats and Microdata, and has shown a significant increase in the use of RDFa and Microdata, and the stagnation in the use of Microformats (by comparing the years 2009/2010 and 2012). The authors also show the use of vocabulary i.e. ontology in the specified formats.

Similar research to [10] conducted [11] by using data from the Internet's search engine *Bing* indexed in January 2012. The same authors conducted a similar research using data of search engine *Yahoo*. Their research differs from the previous one, except in the data source and methods of data extraction and display of the results in the form of triplets. Thus, for example, they note that *Facebook.com* has the most triples if we only look at RDFa, MySpace.com in microdata technology and Yahoo.com in microformats. Besides the number of triplets, research shows the number of pages that contain certain formats separating at the same time RDFa from OGP's[5]. Also it describes the use of ontology for each format. Their analysis showed that 30% of websites contain some form of metadata.

Our research differ from previously mentioned in focus on e-commerce web sites i.e. Internet commerce and the data source itself - we will take into account the top of 100 most popular (according to the traffic) Internet shops (stores). Research will show the current situation in this industry.

Authors of [2] in their work have presented a platform for product search supported by semantic web. This work is a direct example of the use of semantic tagging products in Internet stores. Their solution is available on *www.XploreProducts.com* where visitors can try out the product search. Platform as a source uses its own database that is constantly updated by pinging web shops that use RDFa for semantic labeling. Limitations of this study and the proposed algorithms can be seen in just one technology (RDFa) and ontology (schema.org) that were used. The authors propose further research using eClassOWL and GoodRelations ontology. This article provides a motivation for our research that would show how much potential currently have such systems. To what extent is semantic tagging today present in web stores to make such a system become a reality?

IV. The Research Methodology

To determine the degree of use of semantic annotations in the most popular web shops we used a list of the top 100 Web stores ranked by *Alexa.com* in the category „Shopping".[6] By simple inspection of the HTML website code of product details page we can determine the usage of semantic annotations and vocabularies. For each web shop from the list, we chose a random product and investigate the product page by inserted the URL into the „Google structured data testing tool"[7]which detects structural information on the website. The data obtained are sorted by technology, vocabulary and namespaces.

V. The Results and Analysis

The research results presented in Table I and Table II show the use of certain formats and namespaces or types of entities. The most commonly used is Open Graph Protocol (OGP), which uses 66 web shops, and usually in the most rudimentary form, using only the basic elements. OGP cannot be considered the right format for labeling of products because it is stated in meta data in the header of the page. However, we have included it in our research because it still includes some metadata that serves social networks to display shared link. Purpose of the product description contained in the OGP markings is different from the ordinary product descriptions - OGP descriptions are mostly "call to action" descriptions that are adjusted for users of social networks. Most of the web shops are using them in that way. Only 18 web shops use Microformats in comparison to Microdata that use 44 web shops, which are understandable, since it is the most recent recommendation of major search engines. Surprisingly, there is no use of any other vocabulary or ontology except

[3] http://commoncrawl.org
[4] http://webdatacommons.org
[5] Facebook Open Graph Protocol, http://ogp.me
[6] http://www.alexa.com/topsites/category/Top/Shopping

Schema.org. Also the namespaces that are used are reduced to three main groups: "product", "offer" and "aggreagateRating." More detailed use of other elements is missing. It is also surprising the fact that none of the web shops did use RDFa to indicate structural information about the products. RDFa is used only for OGP in the header. None of the studied web shops use GoodRelations ontology.

TABLE I: FORMATS OF SEMANTIC ANNOTATION OF PRODUCTS USED BY TOP 100 WEB SHOPS

Format	Number of web shops
RDFa (OGP)	66
Microdata	44
Microformats	18

TABLE II: FORMATS AND NAMESPACES (VOCABULARY) USED BY TOP 100 WEB SHOPS

Format	Namespace	Number of web shops
RDFa	opengraphprotocol.org/schema	66
Microdata	schema.org/product	40
	schema.org/offer	30
	schema.org/AggregateRating	23
	schema.org/organization	7
	schema.org/review	8
	schema.org/rating	7
	schema.org/bredcrumb	6
	other	10
Microformat	hreview-aggregate	12
	hproduct	10
	hreview	2

Similar research on a larger corpus of data has shown an upward trend in the use of RDFa and Microdata, while stagnation of Microformats comparing the year 2010 and 2012 [10]. It is obvious that *Schema.org* since the year 2011 has greatly influenced the use of formats for labeling which led to these results, at least in the domain of e-commerce. We can conclude that era of Microformats slowly passes and Microdata markups are increasingly spreading.

Limitation of this research is reflected in the use of relatively small source of data (product pages of top 100 web shops) and no possibility of comparison with previous researches in the domain of e-commerce. Remains unexplored to what extent the formats for labeling of products were used in the corpuses from the 2009/10 and the year 2012 analyzed by [10]. It is not known to what extent are certain e-commerce ontologisms used in *Linked data*, all of which represent a challenge for future research.

APPENDIX

APPENDIX I: TOP 100 WEB SHOPS USED IN RESEARCH

1	amazon.com	51	forever21.com
2	ebay.com	52	wiley.com
3	netflix.com	53	hsn.com
4	amazon.co.uk	54	jcrew.com
5	walmart.com	55	yoox.com
6	Ikea.com	56	sephora.com
7	bestbuy.com	57	pixmania.com
8	target.com	58	cabelas.com
9	groupon.com	59	shopbop.com
10	homedepot.com	60	officedepot.com
11	bodybuilding.com	61	iherb.com
12	newegg.com	62	neimanmarcus.com
13	lowes.com	63	net-a-porter.com
14	macys.com	64	rei.com
15	sears.com	65	focalprice.com
16	gap.com	66	marksandspencer.com
17	autos.yahoo.com	67	urbanoutfitters.com
18	sky.com	68	cambridge.org
19	nordstrom.com	69	fineartamerica.com
20	zappos.com	70	redbubble.com
21	dx.com	71	saksfifthavenue.com
22	overstock.com	72	weightwatchers.com
23	staples.com	73	modcloth.com
24	barnesandnoble.com	74	abebooks.com
25	bhphotovideo.com	75	futureshop.ca
26	nike.com	76	drugstore.com
27	hm.com	77	bloomingdales.com
28	kohls.com	78	potterybarn.com
29	ticketmaster.com	79	vitacost.com
30	souq.com	80	mapsofindia
31	costco.com	81	dickssportinggoods.com
32	autotrader.com	82	musiciansfriend.com
33	shutterfly.com	83	uk.moo.com
34	legacy.com	84	frys.com
35	livingsocial.com	85	officemax.com
36	tigerdirect.com	86	revolveclothing.com
37	cars.com	87	play.com
38	walgreens.com	88	scholastic.com
39	6pm.com	89	adorama.com
40	jcpenney.com	90	landsend.com
41	cvs.com	91	anthropologie.com
42	directtv.com	92	cduniverse.com
43	sony.com	93	petco.com
44	gamestop.com	94	micromaxinfo.com
45	victoriassecret.com	95	harborfreight.com
46	bedbathandbeyond.com	96	ae.com
47	trademe.co.nz	97	createandbarrel.com
48	cafepress.com	98	dish.com
49	stubhub.com	99	cargurus.com
50	samsclub.com	100	dsw.com

[7] http://www.google.com/webmasters/tools/richsnippets

REFERENCES

[1] B. J. Corbitt, T. Thanasankit, and H. Yi, "Trust and e-commerce: a study of consumer perceptions," *Electronic Commerce Research and Applications*, vol. 2, no. 3, pp. 203-215, 2003.

[2] D. Vandic, J. V Dam, and F. Frasnicar, "Facted product search powered by the semantic web," *Decision Support Systems*, vol. 53, pp. 425-437, June 2012.

[3] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific American*, no. 284.5, pp. 28-37, 2001.

[4] M. M. Taye, "Understanding semantic web and ontologies: theory and applications," *Journal of Computing*, vol. 2, no. 6, pp. 182-192, June 2010.

[5] S. Pohorec, M. Zorman, and M. Kokol, "Analysis of approaches to structured data on the web," *Computer Standards & Interfaces*, vol. 36, pp. 256-262, November 2013.

[6] R. Khare, "Microformats: the next (small) thing on the semantic web?" *Internet Computing, IEEE*, vol. 10, no. 1, pp. 68-75, 2006.

[7] J. Ronallo, "HTML5 microdata and schema.org," *Code4Lib Journal*, no.16, February 2012.

[8] B. Adida, M. Birbeck, S. McCarron, and S. Pemberton, *RDFa in XHTML: Syntax and Processing*, W3C Recommendation, 2008.

[9] M. Hepp, "GoodRelations: an ontology for describing products and services offers on the web," in *Proc. the 16th International Conference*

on Knowledge Engineering: Practice and Patterns, Acitrezza, 2008, vol. 5268, pp. 329-346.
[10] H. Mühleisen and C. Bizer, "Web data commons – extracting structured data from two large web corpora," in *Proc. CEUR Workshop*, Lyon, 2012, vol. 937.
[11] P. Mika and T. Potter, "Metadata statistics for a large web corpus," in *Proc. CEUR Workshop*, Lyon, 2012, vol. 937.

Goran Matošević was born in 1977 in Pula, Croatia. He graduated from the Faculty of Economics in Rijeka in the year of 2003 and he finished his master's degree at the Faculty of Economics in Zagreb in the year of 2009, the direction of IT management. Currently he is enrolled in a doctoral program in information science at the Faculty of Organization and Informatics in Varaždin.

He works at the Faculty of Economics and Tourism, at the University of Pula as a research assistant. Previously, he was employed as a web developer in several ICT companies. His areas of interests include the semantic web, information retrieval, web technologies and e-commerce.