Assignment 2 - Social Web

Arthur-Ervin Avramiea
2517642
a.e.avramiea@student.vu.nl

Mihnea Dobrescu-Balaur
2549278
mihnea@linux.com

Zilvinas Kucinskas
zks300
zil.kucinskas@gmail.com

1. THE WEB AND SEMANTIC MARKUP

Much of a web developer's effort is geared towards providing the users with a rich, intuitive interface, that makes the information easy to find. However, most of the users' access to information on the web is nowadays mediated by search engines. The user enters a query and expects the search engine to provide him with websites that serve the content he is interested in. The search engine looks up for the string of text within the indexed database that resulted from web crawling. Nonetheless, textual representation of the information may differ between the different websites, and as and as such a lot of relevant data sources may not be identified. A solution to this issue is adding structure to the website content by annotating the text with metadata, in the lines of commonly accepted standards, which the search engine will be able to use to identify specific bits of data. This metadata tags the text that is to be presented to the user with its semantics, or meaning.In the end, the web developer's interest in making the information readable and easy to access for the user translates into an effort in making the information readable and easy to access for the search engines.

Several standards have been in use for the past years, among which RDFa, Microdata and Microformat[**?**]. Semantic markup technologies are adopted rapidly and on a wide scale, to the extent that, in 2013, 50% of the most popular websites according to Alexa, embed some form of structured data.

Microformats[1] are a way of tagging web pages with semantic markup in order to facilitate the automatic processing of data on the Web. They reuse the existing (X)HTML tags in order to have a low barrier of adoption. The markup is done by using specific keywords in the class attribute of elements, for example by adding the "hproduct" class to a div, it is stated that the contents of the div are data about a product[2]. Then, within that div there can be a mention of the product's brand, and that will have a class "brand".

By using these classes in a structured way, the markup of the page gets increased semantic value and it becomes easier to index and link by search engines. Because of the semantics of a pages content, a search engine can figure out semantic links between the given page and other pages on the Internet, which leads to more relevant search results, which in turn leads to more page views.

To get an idea of how microformats-tagged data looks, we will show an example. Consider this snippet from a web page, representing an address:

```
<p>Claude Debussylaan 34, 1082 MD Amsterdam</p>
```

Using microformats, it's easy to add semantic markup:

```
<p class="adr">
<span class="street-address">Claude Debussylaan 34
</span>,
<span class="postal-code">1082 MD</span>
<span class="locality">Amsterdam</span>
</p>
```

Besides products and addresses, there are other microformats classes: calendar, media, news, recipe, resume, review, contact information and others. They are all supported by modern search engines and publishers can test their semantic markup with tools like the Google Structured Data Testing Tool[3].

Microdata is a WHATWG HTML specification used to markup content of the web page and is recognised by such search providers as Google, Bing and Yahoo

[1] http://microformats.org/

[2] http://microformats.org/wiki/hproduct

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

[4]. It allows to specify HTML elements with machine-readable tags, so machines can understand and interpret the data [5]. One example of markup vacabulary is schema.org [6]. It provides ways to specify persons, events, offers, products, organizations, reviews, multimedia and even more [7]. Google Structured Data Testing Tool [8] supports validation of schema.org markup.

Consider the following snippet example from HTML document:

 Postal address: Eerste Ringdidjkstraat 430, Amsterdam 1097BC, The Netherlands

Using schema.org it could be possible to mark it in the following way:

<p itemprop="address" itemscope
itemtype="http://schema.org/PostalAddress">
Postal address:
Eerste Ringdidjkstraat 430,
Amsterdam
1097BC,
The Netherlands

Usage of demonstrated markup can significantly improve work performed by web crawlers and search engines and increase site traffic, since vocabulary is quite popular and widely used. [9]

2. WEBSITES

2.1 Markup for personal webpage

The dynamics of human interaction has obviously changed with the advent of the social web. And in the hectic rythm of our economy, we need to contact people, or find information about them more often. To be visible on the web, we need a form of presenting ourselves to the others, in the form of a social network account or personal website. It is useful, in this context, to structure the page(s) which contain contact details or other information which may be in the interest of somebody who searches for us, so that the search engine "knows" which information represents our telephone number, job position and company, etc. Not only does this help the search engine to point to such a page when somebody searches for a person, but it also helps the search engine website to provide with a summary of the individuals' information in the results, that is extracted by taking into account the specified metadata. This makes life easier for the search engine user - first of all, if the information which he looks for is already in the search engine, he does not need anymore to open the website and look for the information, second, because it allows the user to discriminate between different persons with the same name, before actually opening the web page.

Within these lines, we have chosen to annotate with semantic markup the personal website of Martin Fowler[10], who has a diverse experience in the software industry, ranging from software engineer to training, authoring and consulting. The web page provides with rich, albeit unstructured information about him. We chose to use the Microdata standard with Schema.org due to the rich semantics of the available schemas, that are capable of representing a wide variety of aspects of a person. One of the facilities of Schema.org, of which we took advantage in our markup, is the possibility to create standalone items for organizations with which the person is affiliated, persons to which he is related, or items defining contact details, postal address, etc. This items than can be referred as properties of the person, and reused within other items, whenever that is needed.

Altough Schema.org is sufficient for the details we have on Martin Fowler's page, if he would have also specified details about his education, skills, experience, scientific publications we would have better used microformat's hResume and hPerson schemas. These have provisions for most of the attributes that comprise a curriculum vitae. A disadvantage, however, with using microformat instead of schema.org, is that we lose the ability to reuse items that we have defined (for example to state that a person is now an alumni of an university, but they are also working at that university) .

2.2 Markup for product page

The increase in the number of online stores fosters competition between companies and allows the user with multiple choices. However, when searching a product across multiple websites, the user has to open the webpage of the on-line store to check for the product details and characteristics. In this sense, annotating the web page of the product with semantic markup would allow the search engine to preview in the results a summary of the characteristics of the product, and thus reduce the time it takes to find the right product and make a choice on the provider.

Marktplaats.nl[11] is such a website, on which users and companies alike can advertise new or second hand products that they sell. We have chosen to markup the data using microformat, with the hProduct schema for de-

[4]http://www.whatwg.org/specs/web-apps/current-work/multipage/microdata.html
[5]https://support.google.com/webmasters/answer/1211158?hl=en
[6]https://schema.org/
[7]https://support.google.com/webmasters/answer/1211158?hl=en
[8]http://www.google.com/webmasters/tools/richsnippets
[9]
[10]http://martinfowler.com/aboutMe.html
[11]http://www.marktplaats.nl/

scribing the product, and hCard schema for describing the seller of the product.

Our choice of microformat was not as much determined by its advantages in this situation over other standards, as by our willingness to test it and see how it works. By comparison, Schema.org has some attributes of the product in the Product schema which are not present in microformat's hProduct, and may be useful for the products description. These include details about the physical properties of the object, such as color, weight, width, height; as well as aggregate ratings.

Although the product we have annotated is intended to be sold only one time (as it is only one second-hand bike available of that type), when a product is sold to multiple consumers by a company, or when multiple persons buy from a company, it is useful to have a review of the product or company. In this respect, although microformat provides with the hReview schema for annotating the reviews which belongs to the product, the Review schema from Schema.org is by far much more detailed.

Given these arguments, although for our product, the markup with microformat sufficed for describing the product and its seller, depending on the characteristics we would like to expose to the user, we would chose microdata with Schema.org instead.

3. PUBLISHER EVALUATIONS

3.1 Review 1 - Arthur-Ervin Avramiea

For the first review, I will take the role of publisher in the evaluation of the markup of a webpage which presents the list of episodes from a season of Game of Thrones, along with the description and an image for each episode. The markup will allow the user to see some summary information when searching for an episode.

The validator that I have used is the Google Structured Data Testing Tool.

The use of Schema.org in this situation is appropriate as it contains the proper semantics for describing in detail tv episodes and seasons. However, checking with the validator unravels some technical issues. First of all, the property partOfSeason does not belong to Schema.org. Looking at the item to which the property points, which has type TvSeries, I assume the markup creator intended to use the partOfSeries property instead.

Another issue is that the TvSeason only points to the first episode. After this, the episode property is used on the first episode to point to the second episode and so on. However, the Episode schema does not have the episode property. To resolve this problem, the TvSeason should point to all the episodes.

The presentation of an image and summary may seem at first enough for the presentation of an episode in the search engine. However, it may be of greater value for the publisher if he would also provide with information regarding the actors which play in the episode, the rating, reviews and so on. This would make the content richer, more informative, and attract more users to the publisher's website. For this purpose, I think that it would be better to markup the episode page, which contains all the details that are part of the tv season page, and many more.

3.2 Review 2 - Arthur-Ervin Avramiea

For this review, I will take the role of consumer, and evaluate the markup of a page which presents some details about two football players who are twin brothers. The overall theme of the website is the presentation of famous twins.

The validator that I have used is the Google Structured Data Testing Tool.

RDFa, the annotation standard used for this case is capable of specifying the metadata used for this case. Two rdfa-nodes are created for each football player, with the type Person. However, as the validator shows, one of the required fields - fn is not specified for the two nodes. Metadata with error may not be taken into account by the search engineand, as a consumer, I will not able to take advantage of the markup, in the form of a Google information box, for example.

The information encapsulated in the metadata contains the names of the football players, a reference to an image, the country of origin, the family relationship(brothers), and their common interest (soccer). However, as a consumer, I may be interested to see more information in the Google information box - such as, which is the football team at which they are playing now, which is their date of birth, and so on.

3.3 Review 1 - Zilvinas Kucinskas

From the perspective of consumer, first site marks content about "Games of Throne" TV series [12] using schema.org language[13]. It does a great job by extracting relevant data about third season episodes, such as image, name, url, publication date. But according to Google Structure Data Testing Tool[14] it has some mistakes too. Information about TV series is duplicated, http://schema.org/tvseries type is used only with a name and additional information is extracted as a separate node without a relationship to a type. Title of the series is tagged as "Game of Thrones (TV Series 2011âĂŽÃĎÃň)", so as a consumer I wouldn't like to get information provided in brackets

[12]http://www.imdb.com/title/tt0944947/episodes?season=3
[13]https://schema.org/
[14]http://www.google.com/webmasters/tools/richsnippets

as it refers more to a type than to a title. Episodes are also duplicated. Episode number, publication date, name and description are extracted as a separate node, and url is provided as a second name. Relationship with the "Games of Throne" series are provided as a separate node, which has image, url and episode number (provided with the legacy spelling, should be changed to singular form). All in all, provider should fix the relationships and spelling to provide the consumer with more comprehensive results, because for example, search engines could not provide me with correct information, when consumer would query for tv series name providing episode number and name of the episode.

3.4 Review 2 - Zilvinas Kucinskas

From the perspective of publisher, second site marks famous soccer players using RDFa language [15]. The page provides markup for picture, name, family name, city and country and relationship to other player, for example, player A is a sibling of player B. It helps web crawlers by providing comprehensive information about players, so they can understand and interpret it. Algorithms can easily capture relationship and use data provided with them, so it provides better user experience for consumers. Furthermore, provider can appear higher in search engine results.

3.5 Review 1 - Mihnea Dobrescu-Balaur

This is a version with semantic markup of the IMDB page for Game of Thrones (season 3) [16]. The markup uses microdata, as seen on schema.org[17]. This review is from the perspective of the publisher.

Schema.org has support for TV series, seasons and episodes, and the annotated page makes use of them extensively. It also created links between them, so that an indexer can know that an episode is part of a season. All this leads to good information extraction and rich snippets in search results.

It seems that the tagging was done using the Google Markup Helper[18], which allowed the publisher to easily add the tags. However, since this was an automated process, there are a few errors. For example, the div that only contains the title of the TV Series is labelled as âĂIJTVEpisodeâĂİ. Otherwise, all the data is annotated - we can see datePublished for air date, descriptions, episode numbers and such.

This kind of annotation is good from a publisherâĂŹs perspective, since it helps with user acquisition and position in search results. There are a few disadvantages as well - because adding the markup by hand is a tedious process, the publisher had to use an automated tool, and the automated tool made a few mistakes. A good balance would be to have a human do a final review before publishing.

3.6 Review 2 - Mihnea Dobrescu-Balaur

This is a version with semantic markup of a page showing famous twins, namely Frank de Boer and Ronald de Boer. The page uses FOAF annotations[19] to display the relationships between the persons depicted. This review is from the perspective of the consumer.

The present annotations make good use of the FOAF markup. They not only represent the relevant data about the persons (like name, family name, location, sport played and others), but they also express the relationship between the two (siblingOf).

This kind of annotation is good from a consumerâĂŹs perspective, since they can programatically download this website and easily parse out the relevant data and relationships, just by looking at the attributes of the HTML nodes. If the page were not annotated, then the consumer would have needed to scrape the page and manually parse the natural language in order to infer the contained facts. Also, thanks to the markup, when using a search engine, looking for one of the two brothers, the results page can automatically display and link to the brother.

[15] http://www.w3.org/TR/xhtml-rdfa-primer/
[16] http://www.imdb.com/title/tt0944947/episodes?season=3
[17] http://schema.org
[18] http://www.google.com/webmasters/markup-helper/
[19] http://www.foaf-project.org/