# 5. Using semantic internet standards

As most people probably noticed, modern search engines do not only present a list of search results, they also give you a box of specific product offers at various web shops that relate to your search. They can only do this if they can somehow relate your search phrase to other content semantically and that is only possible if the content provider adds descriptive semantics to its product offers. The use of these so-called semantic standards is of special interest for statistics for several reasons:

- Generally speaking, the more structured the web is, the easier it will become to obtain structured content for official statistics.
- For web shops, if semantic standards are used to annotate products with properties such as product name, price, description, colour etc. getting these product details is much easier
- If different web shops use the same semantic notions the comparability of data streams across data sources will improve..

It is therefore crucial to keep an eye on the (slowly) emerging semantic internet standards and, in particular, the acceptance and use of these on websites of our interest. After all, from a data consumer viewpoint, it is not so much a question which standard is better or simpler, what counts is if the standard is used. We want to know before we invest in it. Below, we have a look at some different technologies / standards and try to estimate the importance to improve our tools with native knowledge of these vocabularies for the goal of getting structured data for statistics.

## 5.1 Microformats

Microformats and its successor microformats2 (see microformats.org) were launched around 2005 to provide meaning on the web in HTML in simple way. At that time there were also other alternatives, some even from big players, and the microformats movement seems to have rebelled against these, especially to provide simplicity in an open standard. In short microformats are "designed for humans first and machines second, microformats are a set of simple, open data formats built upon existing and widely adopted standards". An example is the use of the hcard / vcard classname to indicate a reference to a person.

```
<div class="vcard">
 <a class="url fn" href="http://tantek.com/">Tantek Çelik</a>
</div>
```

## 5.2 RDFa

The linked data initiative, implemented in RDFa (Rdfa.info) is, again, an extension to HTML5 to markup concepts like people, places, events, recipes etc. to improve listing in search engines and visibility on the web. This example uses the vocab and property attributes to refer to elements in a common vocabulary:

```
<body vocab="http://purl.org/dc/terms/">
  ...
  <h2 property="title">The Trouble with Bob</h2>
  <p>Date: <span property="created">2011-09-10</span></p>
  ...
  <p>All content on this site is licensed under
```

```
   <a property="http://creativecommons.org/ns#license"
href="http://creativecommons.org/licenses/by/3.0/">
     a Creative Commons License</a>. ©2011 Alice Birpemswick.</p>
</body>
```

## 5.3 Microdata

HTML Microdata is a standard from the W3C consortium developed around 2013. It is a way to nest metadata with existing content on web pages, with the focus of telling search engines what content is actually provided. It is supposed to be simpler than competing standards such as RDFa and microformats. Below is an example:

```
<section itemscope itemtype="http://schema.org/Person">
         Hello, my name is
         <span itemprop="name">John Doe</span>,
         I am a
         <span itemprop="jobTitle">graduate research assistant</span>
         at the
         <span itemprop="affiliation">University of Dreams</span>.
         My friends call me
         <span itemprop="additionalName">Johnny</span>.
         You can visit my homepage at
         <a href="http://www.JohnnyD.com"
itemprop="url">www.JohnnyD.com</a>.
         <section itemprop="address" itemscope
itemtype="http://schema.org/PostalAddress">
                  I live at
                  <span itemprop="streetAddress">1234 Peach Drive</span>,
                  <span itemprop="addressLocality">Warner Robins</span>,
                  <span itemprop="addressRegion">Georgia</span>.
         </section>
</section>
```

## 5.4 JSON-LD

JSON-LD (Javascript Object Notation-Linked data) is a recommendation from the W3C consortium released in 2014. It is a lightweigth Linked Data format that builds on the popular JSON format extensively used on the web and creates a network of standards-based, machine readable data across websites. An example:

```
{
  "@context": "http://json-ld.org/contexts/person.jsonld",
  "@id": "http://dbpedia.org/resource/John_Lennon",
  "name": "John Lennon",
  "born": "1940-10-09",
  "spouse": "http://dbpedia.org/resource/Cynthia_Lennon"
}
```

## 5.5 The Open Graph (OG) protocol

The Open Graph protocol (ogp.me) was originally created by facebook and is used to use web pages to in the so called social graph. It focuses on developer simplicity and its initial version is based on RDFa. Although its application seems to be oriented to the social web more than the commercial web we mention it here as it seems to gain some popularity. The following is an example taken from the open graph website for the movie The Rock on IMDB:

```
<html prefix="og: http://ogp.me/ns#">
<head>
<title>The Rock (1996)</title>
<meta property="og:title" content="The Rock" />
<meta property="og:type" content="video.movie" />
```

```
<meta property="og:url" content="http://www.imdb.com/title/tt0117500/"
/>
<meta property="og:image" content="http://ia.media-
imdb.com/images/rock.jpg" />
...
</head>
...
</html>
```

## 5.6 Structured Data Markup

Structured data markup (developers.google.com/structured-data) is a standard created by Google to facilitate annotating content on websites so that machines can understand it. It supports schema.org (see below) vocabularies. Websites may choose between the formats microformats, RDF-a and JSON-LD for embedding semantics in their web page, but JSON-LD is not supported for all types. At time of writing Google says it is in the process of extending JSON-LD support. In a way structured data markup is a generalisation of the three other standards. An example of a Structured data markup product listing in JSON-LD is:

```
{
  "@context": "http://schema.org/",
  "@type": "Product",
  "name": "Executive Anvil",
  "image": "http://www.example.com/anvil_executive.jpg",
  "description": "Sleeker than ACME's Classic Anvil",
  "mpn": "925872",
  "brand": {
    "@type": "Thing",
    "name": "ACME"
  },
  "aggregateRating": {
    "@type": "AggregateRating",
    "ratingValue": "4.4",
    "reviewCount": "89"
  },
  "offers": {
    "@type": "Offer",
    "priceCurrency": "USD",
    "price": "119.99",
    "priceValidUntil": "2020-11-05",
    "itemCondition": "http://schema.org/UsedCondition",
    "availability": "http://schema.org/InStock",
    "seller": {
      "@type": "Organization",
      "name": "Executive Objects"
    }
  }
}
```

## 5.7 Schema.org

Schema.org is a site, sponsored by Google, Microsoft, Yahoo and Yandex, to create, maintain and promote schemas for structured data on the internet and other digital means. The schemas are supposed to be independent from any of the semantic standards mentioned above and are developed by an open community. It claims that it is being used on over 10 million sites to mark up their website pages. This is not so much a technical standard comparable with the above standards; it is more about standardization on content level. It is referenced multiple times in the example in the previous paragraph, for specifying the

condition and the availability. Schema.org has some interesting definitions for Product, Offer, Organisation, Rating, Person etc.

## 5.8 Acceptance

From this brief tour on semantic technologies and standards we conclude that the playground is a bit tough. There are a few different standards with the same goals, sometimes competing, sometimes working together. To get a feeling of the actual use of these standards in practice we reviewed some of the web shops with respect to these standards. We also observed whether the standard was used on the product page (PP) or on the product list page (PL). This is of specific interest for bulk-scraping. If product details can be found on a product list page there is no need to inspect the product details page which saves a lot of internet traffic. Here are the results:

| website | microformat | RDFa | microdata | JSON-LD | OG | PP | PL |
|---|---|---|---|---|---|---|---|
| wehkamp.nl | yes, v1 (description, price, review) | no | yes (name, image, price, availability, description) | no | no | yes | no |
| hm.com | no | no | no | no | yes | yes | no |
| zalando.nl | no | no | yes (brand, name, price, availability) | no | no | yes | no |
| esprit.nl | no | no | no | no | yes | yes | no |
| missetam.nl | no | no | yes (url, name, price, priceCurrency, breadcrumb) | no | no | yes | no |
| wefashion.nl | no | no | yes (productID, name, url, priceCurrency, price, availability) | no | no | no | yes |
| hugoboss.com | no | no | yes (name, url, image, model, color, price, productID) | no | no | yes | no |
| thesting.com | no | no | no | no | yes | yes | no |
| bartsmit.com | yes, v1 (description, price, review) | no | yes (name, offers, priceCurrency, price, availability, image, rating) | no | no | yes | no |
| bol.com | no | no | yes (image, name, brand, logo, sku, description, offers, price, availability) | no | no | yes | yes |
| mediamarkt.nl | no | no | yes (image, name, brand, description, offers, priceCurrency, price, availability) | no | yes | yes | no |

This brief survey shows a clear pattern in favour of the microdata standard. Eight out of eleven sites use the microdata standard in some way. The other three use Open Graph. Microformats, RDFa and JSON-LD are hardly used.

All sites using microdata support name and price tags, which are of specific interest for us. However usually they do this on the product detail page only, which might complicate things a bit for bulk-scraping. All in all, this is a nice result for our feasibility study. Assuming this is representative for the sites of our interest, we conclude that it is worth looking at the possibilities to extend our current tools with some of these standards, especially the microdata standard..