Web Data Commons – Extracting Structured Data from Two Large Web Corpora

Hannes Mühleisen
Web-based Systems Group
Freie Universität Berlin
Germany
muehleis@inf.fu-berlin.de

Christian Bizer
Web-based Systems Group
Freie Universität Berlin
Germany
christian.bizer@fu-berlin.de

ABSTRACT

More and more websites embed structured data describing for instance products, people, organizations, places, events, resumes, and cooking recipes into their HTML pages using encoding standards such as Microformats, Microdatas and RDFa. The Web Data Commons project extracts all Microformat, Microdata and RDFa data from the Common Crawl web corpus, the largest and most up-to-data web corpus that is currently available to the public, and provides the extracted data for download in the form of RDF-quads. In this paper, we give an overview of the project and present statistics about the popularity of the different encoding standards as well as the kinds of data that are published using each format.

1. INTRODUCTION

In recent years, much work has been invested in transforming the so-called "eyeball" web, where information is presented for visual human perception towards a "Web of Data", where data is produced, consumed and recombined in a more or less formal way. A part of this transformation is the increasing number of websites which embed structured data into their HTML pages using different encoding formats. The most prevalent formats for embedding structured data are Microformats, which use style definitions to annotate HTML text with terms from a fixed set of vocabularies; RDFa, which is used to embed any kind of RDF data into HTML pages; and Microdata, a recent format developed in the context of HTML5.

The embedded data is crawled together with the HTML pages by Google, Microsoft and Yahoo!, which use the the data to enrich their search results. These companies have so far been the only ones capable of providing insights into the amount as well as the types of data that are currently published on the Web using Microformats, RDFa and Microdata. While a previously published study by Yahoo! Research [4] provided many insight, the analyzed web corpus not publicly available. This prohibits further analysis and the figures provided in the study have to be taken at face value.

However, the situation has changed with the advent of the *Common Crawl*. Common Crawl[1] is a non-profit foundation that collects data from web pages using crawler software and publishes this data. So far, the Common Crawl foundation has published two Web corpora, one dating 2009/2010 and one dating February 2012. Together the two corpora contain over 4.5 Billion web pages. Pages are included into the the crawls based on their PageRank score, making these corpora snapshots of the popular part of the web.

The Web Data Commons project has extracted all Microformat, Microdata and RDFa data from the Common Crawl web corpora and provides the extracted data for download in the form of RDF-quads. In this paper, we give an overview of the project and present statistics about the popularity of the different encoding formats as well as the kinds of data that are published using each format.

The remainder of this paper is structured as follows: Section 2 gives an overview of the different formats that are used to embed structured data into HTML pages. Section 3 describes and compares the changes in format popularity over time, while Section 4 discusses the kinds of structured data that are embedded into web pages today. Section 5 presents the extraction framework that was used to process the Common Crawl corpora on the Amazon Compute Cloud. Finally, Section 6 summarizes the findings of this paper.

2. EMBEDDING STRUCTURED DATA

This section summarizes the basics about Microformats, RDFa and Microdata and provides references for further reading.

2.1 Microformats

An early approach for adding structure to HTML pages were *Microformats*[2]. Microformats define of a number of fixed vocabularies to annotate specific things such as people, calendar entries, products etc. within HTML pages. Well known Microformats include *hCalendar* for calendar entries according to RFC2445, *hCard* for people, organizations and places according to RFC2426, *geo* for geographic coordinates, *hListing* for classified ads, *hResume* for resume information, *hReview* for product reviews, *hRecipe* for cooking recipes, *Species* for taxonomic names of species and *XFN* for modeling relationships between humans.

For example, to represent a person within a HTML page using the hCard Microformat, one could use the following markup:

```
<span class="vcard">
 <span class="fn">Jane Doe</span>
</span>
```

In this example, two inert `<span>` elements are used to first create a person description and then define the name of the person described. The main disadvantages of Microformats are their case-by-case syntax and their restriction to a specific set of vocabulary terms.

[1] http://http://commoncrawl.org

[2] http://microformats.org

LDOW2012, April 16, 2012, Lyon, France.

To improve the situation, the newer formats, RDFa and Microdata, provide vocabulary-independent syntaxes and allow terms from arbitrary vocabularies to be used.

2.2 RDFa

RDFa defines a serialization format for embedding RDF data [3] within (X)HTML pages. RDFa provides a vocabulary-agnostic syntax to describe resources, annotate them with literal values, and create links to other resources on other pages using custom HTML attributes. By also providing a reference to the used vocabulary, consuming applications are able to discern annotations. To express information about a person in RDFa, one could write the following markup:

```
<span xmlns:foaf="http://xmlns.com/foaf/0.1/"
  typeof="foaf:Person">
  <span property="foaf:name">Jane Doe</span>
</span>
```

Using RDFa markup, we refer to an external, commonly-used vocabulary and define an URI for the thing we are describing. Using terms from the vocabulary, we then select the "Person" type for the described thing, and annotate it with a name. While requiring more markup than the hCard example above, considerable flexibility is gained. A mayor supporter of RDFa is Facebook, which has based its Open Graph Protocol[3] on the RDFa standard.

2.3 Microdata

While impressive, the graph model underlying RDF was thought to represent entrance barriers for web authors. Therefore, the competing *Microdata* format [2] emerged as part of the HTML5 standardization effort. In many ways, Microdata is very similar to RDFa, it defines a set of new HTML attributes and allows the use of arbitrary vocabularies to create structured data. However, Microdata uses key-value pairs as its underlying data model, which lacks much of the expressiveness of RDF, but at the same time also simplifies usage and processing. Again, our running example of embedded structured data to describe a person is given below:

```
<span itemscope
  itemtype="http://schema.org/Person">
    <span itemprop="name">Jane Doe</span>
</span>
```

We see how the reference to both type and vocabulary document found in RDFa is compressed into a single type definition. Apart from that, the annotations are similar, but without the ability to mix vocabularies as it is the case in RDFa. The Microdata standard gained attention as it was selected as preferred syntax by the *Schema.org* initiative, a joint effort of Google, Bing and Yahoo, which defines a number of vocabularies for common items and carries the promise that data that is represented using these vocabularies will be used within the applications of the founding organizations.

3. FORMAT USAGE

As of February 2012, the Common Crawl foundation have released two Web corpera. The first corpus contains web resources (pages, images, ...) that have been crawled between September 2009 and September 2010. The second corpus contains resources dating February 2012, thereby yielding two distinct data points for which we can compare the usage of structured data within the pages. As a first step, we have filtered the web resources contained in the corpora to only include HTML pages. Table 1 shows a comparison of the two corpora. We can see how HTML pages represent the bulk of the corpora. The newer crawl contains fewer web pages. 148 million HTML pages within the 2009/2010 crawl contained structured data, while 189 million pages within the 2012 crawl contained structured data. Taking the different size of the crawl into account, we can see that the fraction of web pages that contain structured data has increased from 6 % in 2010 to 12 % in 2012. The absolute numbers of web pages that used the different formats are given in Table 2. The data sets that we extracted from the corpora consist of 3.2 billion RDF quads (2012 corpus) and 5.2 billion RDF quads (2009/2010 corpus).

[3]http://ogp.me/

	2009/2010	2012
Crawl Dates	09/09 – 09/11	02/12
Total URLs	2.8B	1.7B
HTML Pages	2.5B	1.5B
Pages with Data	148M	189M

Table 1: Comparison of the Common Crawl corpora

Format	2009/2010	2012
RDFa	14,314,036	67,901,246
Microdata	56,964	26,929,865
geo	5,051,622	2,491,933
hcalendar	2,747,276	1,506,379
hcard	83,583,167	61,360,686
hlisting	1,227,574	197,027
hresume	387,364	20,762
hreview	2,836,701	1,971,870
species	25,158	14,033
hrecipe	115,345	422,289
xfn	37,526,630	26,004,925

Table 2: URLs using the different Formats

Fig. 1 shows the distribution of the different formats as a percentage of the number of URLs with the respective format to the total number of URLs within the corpora and compares the fractions for the 2009/2010 corpus and the 2012 corpus. We see that RDFa and Microformats gain popularity, while the usage of the single-purpose Microformats remain more or less constant. The reason for the explosive adoption of the Microdata syntax between 2010 and 2012 might be announcement in 2011 that Microdata is the preferred syntax of the *Schema.org* initiative.

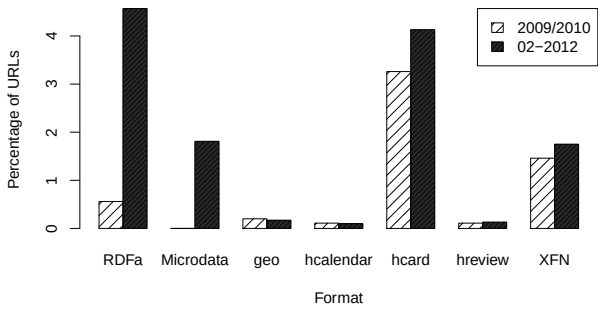

Figure 1: Common Crawl Corpora – Format Distribution

The study by Yahoo! Research [4] confirms our observations. The study is based on a Yahoo corpus consisting of 12 Billion web pages. The analysis was repeated three times between 2008 and 2010 to investigate the development of the formats. They measured the percentage of URLs that contained the respective format. These results are shown in Fig. 2.

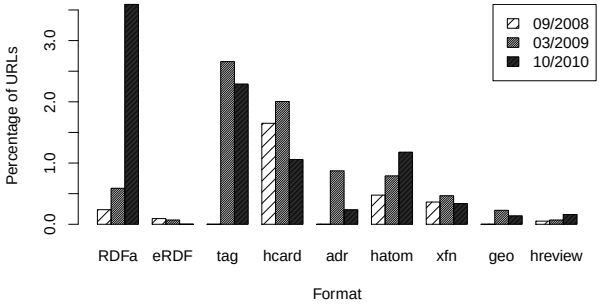

Figure 2: Yahoo! Corpora – Format Distribution (adopted fom [4])

We can see that RDFa exhibits non-linear growth, with the other formats are not showing comparable developments. A second survey on the format distribution was presented by Sindice, a web index specializing in structured data [1]. Their survey was based on a set of 231 Million web documents collected in 2011. Their results were similar to the 2010 sample from the Yahoo! survey, showing major uptake for RDFa.

4. TYPES OF DATA

While each Microformat can only be used to annotate the specific types of data it was designed for, RDFa and Microdata are able to use arbitrary vocabularies. Therefore, the format comparison alone does not yield insight into the types of data being published. RDFa and Microdata both support the definition of a data type for the annotated entities. Thus simple counting the occurrences of these types can give an indicator of their popularity. The top 20 values for type definitions of the RDFa data within the 2012 corpus are given in Table 3. Type definitions are given as shortened URLs, using common prefixes[4]. Note that *gd:* stands for Google's *Data-Vocabulary*, one of the predecessor of *Schema.org*.

We have then manually grouped the 100 most frequently occurring types by entity count into groups. These groups are given in Table 4. The most frequent types were from the area of website structure annotation, where for example navigational aides are marked. The second most popular area are information about people, businesses and organizations in general, followed by media such as audio files, pictures and videos. Product offers and corresponding reviews represent the fourth most frequent group, and geographical information such as addresses and coordinates was least frequent. Groups below 1 % frequency are not given.

Table 6 shows the same analysis for the Microdata data within the 2012 corpus. Apart from variations in the specific percentages, the same groups were found to be most frequently used. An interesting observation was that only two of the 100 most frequently occurring types were *not* from of the Schema.org namespace, confirming the overwhelming prevalence of types from this namespace, which is not surprising since the Microdata format itself was made popular by this initiative. This is also shown in the listing of the 20 most frequent type definitions given in Table 5, where all URLs originate from either the *Schema.org* domain or the *Data-Vocabulary* domain.

[4] `http://prefix.cc/popular.file.ini`

Type	Entities
gd:Breadcrumb	13,541,661
foaf:Image	4,705,292
gd:Organization	3,430,437
foaf:Document	2,732,134
skos:Concept	2,307,455
gd:Review-aggregate	2,166,435
sioc:UserAccount	1,150,720
gd:Rating	1,055,997
gd:Person	880,670
sioctypes:Comment	666,844
gd:Product	619,493
gd:Address	615,930
gd:Review	540,537
mo:Track	444,998
gd:Geo	380,323
mo:Release	238,262
commerce:Business	197,305
sioctypes:BlogPost	177,031
mo:SignalGroup	174,289
mo:ReleaseEvent	139,118

Table 3: Top-20 Types for RDFa

Area	% Entities
Website Structure	29 %
People, Organizations	12 %
Media	11 %
Products, Reviews	10 %
Geodata	2 %

Table 4: Entities by Area for RDFa

To further investigate the lack of diversity that has become apparent in the analysis of the 100 most frequently used types for RDFa and Microdata, we have calculated a histogram for the most frequently used types. These histograms are shown in Fig. 3. For both corpora, the histogram of type frequencies is plotted on a logarithmic scale. From the graph, we can make two observations: A small number of types enjoy very high popularity, and the long tail is rather short. For both formats, no more than 200 types had more than 1000 instances.

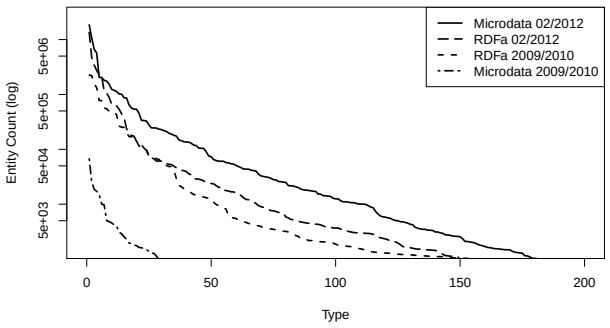

Figure 3: Microformat / RDFa type frequency

Type	Entities
gd:Breadcrumb	18,528,472
schema:VideoObject	10,760,983
schema:Offer	6,608,047
schema:PostalAddress	5,714,201
schema:MusicRecording	2,054,647
schema:AggregateRating	2,035,318
schema:Product	1,811,496
schema:Person	1,746,049
gd:Offer	1,542,498
schema:Article	1,243,972
schema:WebPage	1,189,900
gd:Rating	1,135,718
schema:Review	1,016,285
schema:Organization	1,011,754
schema:Rating	872,688
gd:Organization	861,558
gd:Product	647,419
gd:Person	564,921
gd:Review-aggregate	539,642
gd:Address	538,163

Table 5: Top-20 Types for Microdata

Area	% Entities
Website Structure	23 %
Products, Reviews	19 %
Media	15 %
Geodata	8 %
People, Organizations	7 %

Table 6: Entities by Area for Microdata

5. EXTRACTION PROCESS

The Common Crawl data sets are stored in the AWS Simple Storage Service (S3), hence extraction was also performed in the Amazon cloud (EC2). The main criteria here are the costs to achieve a certain task. Extracting structured data had to be performed in a distributed way in order to finish this task in a reasonable time. Instead of using the ubiquitous Hadoop framework, we found using the Simple Queue Service (SQS) to coordinate for our extraction process increased efficiency. SQS provides a message queue implementation, which we used to co-ordinate 100 extraction nodes.

The Common Crawl corpora were already partitioned into compressed files of around 100MB each. We added the identifiers of each of these files as messages to the queue. The extraction nodes share this queue and take file identifiers from it. The corresponding file was then downloaded from S3 to the node. The compressed archive was split into individual web pages. On each page, we ran our RDF extractor based on the Anything To Triples (Any23) library. The resulting RDF triples were written back to S3 together with extraction statistics and later collected.

Any23 parses web pages for structured data by building a DOM tree and then evaluates XPath expressions to extract the structured data. While profiling, we found this tree generation to account for much of the parsing cost, and we have thus searched for a way to reduce the number of times this tree is built. Our solution was to run regular expressions against each archived web page prior to extraction, which detected the presence of structured data within the HTML page, and only to run the Any23 extractor when the regular expression found potential matches.

The costs for parsing the 28.9 Terabytes of compressed input data of the 2009/2010 Common Crawl corpus, extracting the RDF data and storing the extracted data on S3 totaled 576 EUR (excluding VAT) in Amazon EC2 fees. We used 100 spot instances of type c1.xlarge for the extraction which altogether required 3,537 machine hours. For the 20.9 Terabytes of the February 2012 corpus, 3,007 machine hours at a total cost of 523 EUR were required.

6. CONCLUSION

The analysis of the two Common Crawl corpora has shown that the percentage of web pages that contain structured data has increased from 6 % in 2010 to 12 % in 2012. The analysis showed an increasing uptake of RDFa and Microdata, while the Microformat deplyoment stood more or less constant.

The analysis of the types of the annotated entities revealed that the generic formats are used to annotate web pages with structural information (breadcrumps) as well as to embed data describing people, organizations, media files, e-commerce data such as products and corresponding reviews and geographical information such as coordinates. Further analysis of the usage frequency of the type definitions of the annotated entities showed a very short tail, with less than 200 significant types. The deployed types as well as the deployed formats seem to closely correlate to the announced support of the big web companies for specific types and formats, meaning that Google, Microsoft, Yahoo almost exclusively determine adoption.

We hope that the data we have extracted from the two web crawls will serve as a resource for future analysis, enabling public research on a topic that was previously almost exclusive to organizations with access to large web corpora. More detailed statistics about the extracted data as well as the extracted data itself are available at `http://webdatacommons.org`.

Acknowledgements

We would like to thank the Common Crawl foundation for creating and publishing their crawl corpora as well as the Any23 team for their structured data extraction framework. This work has been supported by the PlanetData and LOD2 projects funded by the European Community's Seventh Framework Programme.

7. REFERENCES

[1] S. Campinas, D. Ceccarelli, T. E. Perry, R. Delbru, K. Balog, and G. Tummarello. The sindice-2011 dataset for entity-oriented search in the web of data. In *EOS, SIGIR 2011 workshop, July 28, Beijing, China*, 2011.

[2] I. Hickson. HTML Microdata. `http://www.w3.org/TR/microdata/`, 2011. Working Draft.

[3] G. Klyne and J. J. Carroll. *Resource Description Framework (RDF): Concepts and Abstract Syntax - W3C Recommendation*, 2004. http://www.w3.org/TR/rdf-concepts/.

[4] P. Mika. Microformats and RDFa deployment across the Web . `http://tripletalk.wordpress.com/2011/01/25/rdfa-deployment-across-the-web/`, 2011.