Deployment of RDFa, Microdata, and Microformats on the Web – A Quantitative Analysis

Christian Bizer[1], Kai Eckert[1], Robert Meusel[1], Hannes Mühleisen[2], Michael Schuhmacher[1], and Johanna Völker[1]

[1] Data and Web Science Group – University of Mannheim, Germany
firstname@informatik.uni-mannheim.de
[2] Database Architectures Group, Centrum Wiskunde & Informatica, Netherlands
hannes@cwi.nl

Abstract. More and more websites embed structured data describing for instance products, reviews, blog posts, people, organizations, events, and cooking recipes into their HTML pages using markup standards such as Microformats, Microdata and RDFa. This development has accelerated in the last two years as major Web companies, such as Google, Facebook, Yahoo!, and Microsoft, have started to use the embedded data within their applications. In this paper, we analyze the adoption of RDFa, Microdata, and Microformats across the Web. Our study is based on a large public Web crawl dating from early 2012 and consisting of 3 billion HTML pages which originate from over 40 million websites. The analysis reveals the deployment of the different markup standards, the main topical areas of the published data as well as the different vocabularies that are used within each topical area to represent data. What distinguishes our work from earlier studies, published by the large Web companies, is that the analyzed crawl as well as the extracted data are publicly available. This allows our findings to be verified and to be used as starting points for further domain-specific investigations as well as for focused information extraction endeavors.

Keywords: Web Science, Web of Data, RDFa, Microdata, Microformats

1 Introduction

In order to support web applications to understand the content of HTML pages an increasing number of websites have started to semantically markup their pages using different markup formats. The most prevalent of these standards are Microformats[3], which use style definitions to annotate HTML text with terms from a fixed set of vocabularies; RDFa [1], which is used to embed any kind of RDF data into HTML pages; and Microdata [6], a recent format developed in the context of HTML5.
The embedded data is crawled together with the HTML pages by search engines, such as Google, Yahoo!, and Bing, which use the data to enrich their search results [5,3]. These companies have also so far been the only ones capable of providing insights [8,9] into the amount as well as the types of data that are published on the Web using Microformats,

[3] http://microformats.org/

RDFa, and Microdata as they were the only ones possessing large-scale Web crawls. However, the situation has changed with the advent of the *Common Crawl*[4]. Common Crawl is a non-profit foundation that crawls the Web and regularly publishes the resulting Web corpora for public usage on Amazon S3[5].

In this paper, we analyze the deployment of RDFa, Microdata, and Microformats based on the latest Web corpus that has been published by the Common Crawl foundation. The paper makes the following contributions: 1. It presents the first integrated study about the adoption of RDFa, Microdata, and Microformats that is based on a large-scale, publicly-accessible Web corpus and is thus scientifically verifiable. 2. We identify the main topical areas of the published data as well as the vocabularies that are commonly used in order to represent data. 3. We give an impression about the structural richness of the published data by analyzing which properties are used to describe popular types of entities as well as by analyzing the co-occurrence relationships between different types on the same website. 4. Our results can serve as a starting point for further domain-specific investigations as well as focused information extraction endeavors, as we provide all extracted data for public download via the `WebDataCommons.org` website.

The remainder of this paper is structured as follows: Section 2 describes the Common Crawl corpus, while Section 3 gives an overview of the data extraction framework that was used to process the corpus. Section 4 summarizes our overall findings concerning the adoption of the different markup standards. After elaborating on our findings concerning the the deployment of RDFa and analyzing the main topical areas of the RDFa data (Section 5), we detail on the deployment of Microdata (Section 6), and Microformats (Section 7). Section 8 compares our results to related work.

2 The Common Crawl Corpus

The analysis presented in this paper is based on the most recent Web crawl provided by Common Crawl foundation. This Web crawl contains 3,005,629,093 unique HTML pages which originate from 40.6 million pay-level-domains (PLDs). The corpus was crawled in the time span between January 27, 2012 and June 05, 2012. The size of the corpus in compressed form is 48 terabyte. The crawler that is used by the Common Crawl foundation for gathering the corpus relies on the PageRank algorithm for deciding which pages to retrieve. This makes the Common Crawl corpus a snapshot of the popular part of the Web. On the other hand, it also results in the number of pages that are crawled per website to vary widely. For instance, *youtube.com* is represented by 93.1 million pages within the crawl, whereas 37.5 million PLDs are represented by less than 100 pages.

3 The Data Extraction Process

The Common Crawl corpus is published in the form of ARC files which can be obtained from Amazon S3[6]. In order to extract RDFa, Microdata, and Microformats data from

[4] `http://commoncrawl.org`
[5] `http://aws.amazon.com/datasets/41740`
[6] `s3://aws-publicdatasets/common-crawl/parse-output/`

the corpus, we developed a parsing framework which can be executed on *Amazon EC2* and supports the parallel extraction from multiple ARC files. The framework relies on the *Anything To Triples (Any23)*[7] parser library for extracting RDFa, Microdata, and Microformats from the corpus. Any23 outputs RDF quads, consisting of subject, predicate, object, and a URL which identifies the HTML page from which the triple was extracted. For processing the Common Crawl corpus on Amazon EC2, we used 100 AWS x1.large machines. Altogether, extracting the HTML-embedded data from the corpus required 5,636 machine hours amounting to a total machine rental fee of $398.72 using Amazon spot instances. As the number of pages that are contained in the Common Crawl from a single pay-level-domain varies widely, most of the analysis presented in the following is performed using statistics that are aggregated per PLD. In order to determine the PLD of a HTML page, we used the *Public Suffix List*[8]. Hence, a PLD not always equals the second level domain, but country specific domains such as *co.uk* or mass hosting domains like *appspot.com* are considered as top level domains in our experiments. We used *Apache Pig*[9] on Amazon to aggregate the extracted data into a PLD-class-property matrix for each format. We used *Rapidminer*[10] for the vocabulary term co-occurrence analyses that will be presented in the following. The generated RDF dataset as well as the PLD-class-property matrixes are provided for download on the Web Data Commons (WDC) website[11].

4 Overall Results

This section reports our findings concerning the overall deployment of the different markup formats. We discovered structured data within 369 million out of the 3 billion pages contained in the Common Crawl corpus (12.3%). The pages containing structured data originate from 2.29 million among the 40.6 million websites (PLDs) contained in the corpus (5.64%). The RDF representation of the extracted data consists of 7.3 billion RDF quads, describing around 1.15 billion typed entities.

Deployment by Format: Table 1 shows the overall deployment of the three different formats. The second column contains the absolute number of websites that use a specific format. The third column sets these numbers in relation to the overall number of websites covered by the Common Crawl (40.6 million). In column 4, the number of pages containing the respective format is provided. In addition, the table lists the number of typed entities and triples we extracted from the pages containing structured data. Approximately 519 thousand websites use RDFa, while only 140 thousand websites use Microdata. Microformats are used on 1.7 million websites. It is interesting to see that Microformats are used by approximately 2.5 times as many websites as RDFa and Microdata together, despite of the usage of RDFa and Microdata currently being propagated by the major search engines and social networking platforms.

[7] http://any23.apache.org/
[8] http://publicsuffix.org/list/
[9] http://pig.apache.org/
[10] http://rapid-i.com/content/view/181/
[11] http://webdatacommons.org/2012-08/index.html

	#PLDs	%PLDs	#URLs	%URLs	#Typed Entities	#Triples
RDFa	519,379	1.28	169m	5.61	188m	1.01b
Microdata	140,312	0.35	97m	3.23	266m	1.49b
Microformats	1,798,782	4.45	158m	5.26	699m	4.78b

Table 1. Distribution of deployment across the 3 different formats.

Deployment by Popularity of Website: Alexa Internet Inc. maintains a list of the most frequently visited websites. In order to find out how many of the most popular websites provide structured data, we analyzed the deployment of RDFa, Microdata and Microformats on websites that are in the Alexa list of the top 1 million websites[12]. The results of our analysis are given in the four rightmost columns of Table 2 and show that the percentage of the Alexa-listed websites providing structured data (74.75% of the top 100 and 20.56% of the top 1 million) is significantly higher than the percentage of all websites within the Common Crawl that contain structured data (5.64%).

First x	PLDs in CC		% containing structured data			
in AL	#	% AL	overall	RDFa	Microdata	Microformats
100	99	99.00	74.75	34.34	55.56	68.69
1k	963	96.30	62.62	40.08	31.67	46.11
10k	9,294	92.94	47.34	30.47	15.55	29.75
100k	85,058	85.01	31.94	16.46	7.20	20.07
1m	734,882	73.49	20.56	7.55	3.04	14.18

Table 2. Coverage of the PLDs in the Alexa top 1 million list (AL) by the Common Crawl corpus and percentage of these PLDs containing structured data.

Deployment by Top-Level-Domain: Table 3 lists the distribution of websites in the Common Crawl corpus by top-level-domains (TLDs). The last two columns show the number and percentage of the websites by TLD that embed structured data. We see that structured data is provided within all TLDs. In general, the deployment is stronger within generic TLDs like *com* and *net* compared to the country specific TLDs.

Deployment of Multiple Formats on the same Website: As websites could decide to use multiple formats in parallel in order to make it easier for applications to understand their data, we also analyzed the joint usage of two or more formats on the same website. 93,5% of all websites which include structured data use only a single format. 3.7% of the websites contain RDFa alongside with Microformats, while only 1.5% use Microdata together with Microformats. Less than 1% of the websites use Microformats together with RDFa, or all three formats together.

In the following, we discuss the deployment of RDFa, Microdata and Microformats in more detail.

[12] `http://www.alexa.com/topsites` as of Oct 31, 2012.

	TLD	#PLDs in CC	PLDs providing structured data #	%		TLD	#PLDs in CC	PLDs providing structured data #	%
1	com	19,950,689	1,317,757	6.61	11	com.au	428,164	21,400	5.00
2	de	2,810,040	79,366	2.82	12	fr	425,204	29,794	7.01
3	net	2,203,474	145,547	6.61	13	ch	390,336	8,659	2.22
4	org	2,064,960	152,977	7.41	14	pl	382,670	15,524	4.06
5	co.uk	1,448,245	64,043	4.42	15	cz	368,429	11,271	3.06
6	nl	951,484	28,820	3.03	16	ca	319,055	20,938	6.56
7	ru	699,275	27,496	3.93	17	jp	288,267	14,248	4.94
8	info	663,451	46,633	7.03	18	se	286,740	16,649	5.81
9	it	620,726	20,068	3.23	19	eu	259,105	11,429	4.41
10	com.br	501,720	20,418	4.07	20	dk	247,693	10,766	4.35

Table 3. Absolute and relative occurrence of structured data within the top 20 TLDs, ordered by PLD count within the Common Crawl.

5 RDFa Deployment

We discovered 519,379 websites that contain RDFa data, which means that 21% of all websites that contain structured data use RDFa. The share of the websites that use RDFa and belong to the Alexa 1 Million list is 7.55% (see Table 2). Examples of websites from the Alexa top 100 list that use RDFa are the Internet Movie Database (IMDb), the Microsoft news portal and also the website of the British Broadcasting Corporation.

Class/Property Frequency Distribution: In order to determine the topical areas of the published data, we analyzed the vocabularies that are used together with RDFa. Altogether, we discovered that only 98 different classes and 271 properties are used by at least 100 different websites each. The class and property frequency distribution is given in Fig. 1. The x-axis shows the classes and properties, ordered descending by the number of websites that use them. The website count is plotted on the log-scaled y-axis. The frequency of both, classes and properties, follows a long-tailed distribution, i.e. a small number of classes/properties is used very frequently, while the remaining classes/properties are used much less frequent.

Frequent Classes: In order to give an overview of the topical areas of the published data, we analyzed how many websites use specific classes. Table 4 lists the most frequently used RDFa classes together with the number of websites using each class. The namespaces of the classes are abbreviated with the corresponding prefix from the prefix.cc list. In addition to the absolute usage count, the third column in the table shows the relative class usage compared to all websites that embed RDFa. The 4th and 5th column show the usage of RDFa on websites that are contained in the Alexa 1 million list. We see that 6 of the most frequently used classes belong to the Open Graph Protocol (prefix: *og*), and thus to the Facebook ecosystem[13]. In addition we find classes which belong to the topical area of e-commerce (products, offers, reviews, companies) as well as blogging (blog, blogposts, comments). In the following, we discuss these areas in more detail.

[13] `https://developers.facebook.com/docs/concepts/opengraph/`

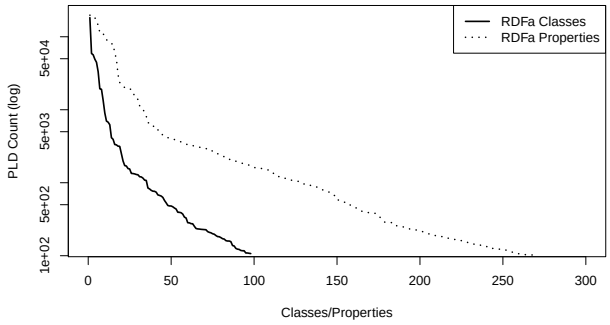

Fig. 1. RDFa class and property distribution by PLD count.

	PLDs Total		PLDs in Alexa			PLDs Total		PLDs in Alexa	
Class	#	%	#	%	Class	#	%	#	%
1 *og:"article"*	183,046	35.24	17,002	30.29	13 *dv:Review -aggregate*	6,236	1.20	1,410	2.51
2 *og:"blog"*	58,971	11.35	5,820	10.37					
3 *og:"website"*	56,573	10.89	9,533	16.98	14 *dv:Rating*	4,139	0.80	845	1.51
4 *foaf:Document*	49,252	9.48	2,802	4.99	15 *sioct:BlogPost*	3,936	0.76	308	0.55
5 *foaf:Image*	44,644	8.60	2,794	4.98	16 *sioct:Comment*	3,339	0.64	456	0.81
6 *sioc:Item*	33,141	6.38	2,188	3.90	17 *og:"activity"*	3,303	0.64	606	1.08
7 *sioc:UserAccount*	19,331	3.72	1,327	2.36	18 *vcard:Address*	3,167	0.61	401	0.71
8 *og:"product"*	19,107	3.68	3,389	6.04	19 *gr:BusinessEntity*	3,155	0.61	392	0.70
9 *skos:Concept*	13,477	2.59	1,135	2.02	20 *dv:Organization*	2,502	0.48	367	0.65
10 *dv:Breadcrumb*	9,054	1.74	2,123	3.78	.. ..	..	..	..	..
11 *sioc:Post*	6,994	1.35	691	1.23	25 *dv:Product*	1,544	0.30	185	0.33
12 *og:"company"*	6,758	1.30	1,067	1.90	26 *gr:Offering*	1,342	0.26	290	0.52

Table 4. Most frequently used RDFa classes.

Facebook Data: The Open Graph Protocol (OGP) is developed and promoted by Facebook in order to enable the integration of external content into the social networking platform. In contrast to RDFa, OGP allows the usage of literals instead of URIs to identify classes. For this reason we mark the names of OGP classes with quotes in Table 4. The 3 most frequently used RDFa classes are the OGP classes *og:"article"*, *og:"blog"* and *og:"website"*. In order to give an indication about the richness of the OGP data, Table 5 shows the properties that are most frequently used together with the top 4 OGP classes. As we can see, the frequent properties are rather generic and the same properties are used for all 4 classes. We see that the old OGP namespace *ogo:* that was officially replaced by the new namespace *ogm:* in the mid of July 2010[14] is still more frequently used than the new one (classes *og:"article"*, *og:"blog"*, and *og:"product"*). We can also observe that the OGP properties are not mixed with properties from other

[14] The namespace *opengraphprotocol.org* was replaced by *ogp.me*. `http://web.archive.org/web/20100719042423/http://opengraphprotocol.org/`

non-Facebook-related vocabularies. Sites using one of the 3 OGP classes *og:"article"*, *og:"blog"* and *og:"website"* use on average 10,08 different properties (at least once).

	OGP class							
	og:"article"		*og:"blog"*		*og:"website"*		*og:"product"*	
Property	#	%	#	%	#	%	#	%
ogo:type	146,836	80.22	42,236	71.62	25,601	45.25	12,263	64.18
ogo:title	142,648	77.93	37,767	64.04	25,043	44.27	12,154	63.61
ogo:url	142,226	77.70	39,201	66.48	24,630	43.54	11,867	62.11
ogo:site_name	126,280	68.99	42,016	71.25	23,524	41.58	11,447	59.91
ogo:description	111,873	61.12	20,131	34.14	21,195	37.46	10,696	55.98
ogo:image	109,283	59.70	19,929	33.79	19,212	33.96	12,008	62.85
fb:app_id	48,403	26.44	29,222	49.55	13,533	23.92	4,241	22.20
ogm:type	36,716	20.06	16,022	27.17	31,411	55.52	6,539	34.22
fb:admins	36,600	19.99	25,900	43.92	17,445	30.84	5,403	28.28
ogm:title	36,349	19.86	15,355	26.04	30,333	53.62	6,466	33.84
ogm:url	35,519	19.40	15,282	25.91	30,423	53.78	6,253	32.73
ogm:site_name	34,173	18.67	15,870	26.91	26,115	46.16	5,892	30.84
ogm:description	30,209	16.50	10,310	17.48	25,572	45.20	5,426	28.40
ogm:image	27,587	15.07	10,068	17.07	24,240	42.85	5,897	30.86

Table 5. Absolute and relative usage of the top properties co-occurring with all the 4 most frequently used OGP classes, ordered by usage frequency with *og:"article"*.

Product Data: We identified three RDFa classes describing products, *og:"product"*, *dv:Product*, and *gr:Offering*, to occur on at least 500 different websites.
The most frequently employed class is *og:"product"* which is used by 19,107 websites (cf. Table 4). The two other product-related classes, *gr:Offering* and *dv:Product*, appear about 10 times less often than *og:"product"* with only 1,544 and 1,342 websites, respectively. The *www.data-vocabulary.org/* vocabulary (*dv:*) was introduced by Google and is declared deprecated since June 2011 in favour of the *schema.org* vocabulary. *gr:Offering* belongs to the GoodRelations vocabulary, an expressive vocabulary for representing e-commerce related data. Analyzing the co-occurrence of *gr:Offering* with other classes from the GoodRelations vocabulary, we found that *gr:Offering* co-occurs in 80.25% of the websites together with *gr:BusinessEntity*. Furthermore, 54.92% of the 1,544 websites also contain *gr:UnitPriceSpecification* in addition to these two classes. The websites that employ *gr:Offering* use on average 27.68 different properties, while websites employing *og:"product"* only use 10.3 different properties to markup their content.

Blog and Document Metadata: The list of the most frequently used RDFa classes given in Table 4 contains 6 classes for annotating individual blog posts, comments, and other article-like web content that is likely published with the help of a content management system: *og:"article"*, *foaf:Document*, *sioc:Item*, *sioc:Post*, *sioct:BlogPost*, and *sioc:Comment*. As for products, we see a dominance of the Open Graph Protocol as *og:"article"* is used by 183,046 websites. Of the 49,252 websites using the *foaf:Document* class, 66% also use *sioc:Item* . From the 33,141 websites using *sioc:Item*, 99% also use of the *foaf:Document* class. Other *sioc:* classes did not show a comparable high

co-occurrence ratio. A possible explanation of the high co-occurrence between *sioc:Item* and *foaf:Document* could be the Drupal 7 CMS. Drupal 7 is a widely used web content management system which supports RDFa natively and marks every page per default as both, a *sioc:Item* and a *foaf:Document*.[15]

Dublin Core: The *dc:* vocabulary is designed to represent metadata describing documents. We found RDFa encoded Dublin Core metadata to be provided by 63,668 websites within the Common Crawl corpus. The most commonly used Dublin Core property is *dc:title*. It was found in 59,957 websites which equals 94.17% of all websites using the *dc:* vocabulary. The second most employed property *dc:date* is embedded in only 20,768 websites.

Creative Commons: The *cc:* vocabulary defines properties for representing licensing information about Web content. The license information is for example used by the Google Image Search to create filters for specific image usage rights. The Creative Commons vocabulary is used by 22,130 websites within the Common Crawl corpus. The two most frequent properties are *cc:attributionURL* and *cc:attributionName* which can be found on 20,195 respectively 20,069 different websites. Following up is *cc:morePermissions* which is used by 4,158 websites.

6 Microdata Deployment

We found 140,312 websites that use Microdata (see Table 1), which means that 6.1% of all websites including structured data use Microdata. The share of the websites that use Microdata and belong to the Alexa Top 1000 list is 31.67% (see Table 2), meaning that Microdata is more widely used by popular websites. Examples of websites from the Alexa Top 100 list that use Microdata are the auction site eBay as well as the websites of Microsoft Corporation and Apple Inc.

Class/Property Frequency Distribution: The frequency distribution of the Microdata classes and properties is given in Fig. 2. The figure shows that only 86 classes and 487 properties are used on more than 100 websites. While the class count is similar to the class count reported for RDFa (98 classes) in Section 5, the number of Microdata properties used is about twice as large as the number of RDFa properties indicating that Microdata annotations are on average more fine grained than RDFa annotations. Regarding the website counts, we observe a long-tailed distribution for classes and properties which is similar to the RDFa distribution (see Fig. 1).

Frequent Classes: Table 6 shows the most frequently used Microdata classes. In addition to the absolute usage count, the third column in the table shows relative class usage compared to all websites having embedded Microdata information. The 4th and 5th column show the usage of Microdata on websites that are contained in the Alexa 1 million list. We see that all frequently used classes either belong to the *schema.org* vocabulary or the *data-vocabulary.org* vocabulary (*dv:*), that was declared deprecated in 2011 in favor of *schema.org*. No classes from any other vocabulary are used together with the Microdata syntax on more than 100 websites. In the following, we discuss the main topical areas of the data.

[15] `http://groups.drupal.org/node/22231`.

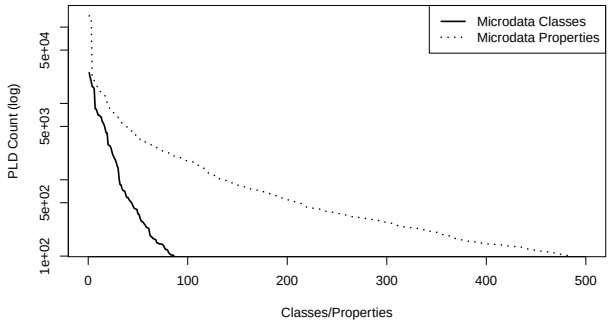

Fig. 2. Microdata class and property distribution by PLD count

		PLDs Total		PLDs in Alexa				PLDs Total		PLDs in Alexa	
	Class	#	%	#	%		Class	#	%	#	%
1	*s:BlogPosting*	25,235	17.98	1,502	6.63	14	*d:Organization*	5,853	4.17	654	2.89
2	*d:Breadcrumb*	21,729	15.49	5,244	23.13	15	*d:Address*	5,559	3.96	654	2.89
3	*s:PostalAddress*	19,592	13.96	1,404	6.19	16	*s:Person*	5,237	3.73	890	3.93
4	*s:Product*	16,612	11.84	3,038	13.40	17	*s:GeoCoordinates*	4,677	3.33	312	1.38
5	*s:LocalBusiness*	16,383	11.68	845	3.73	18	*s:Place*	4,131	2.94	488	2.15
6	*s:Article*	15,718	11.20	3,025	13.35	19	*s:Event*	4,102	2.92	659	2.91
7	*d:Review-aggregate*	8,517	6.07	2,376	10.48	20	*d:Person*	2,877	2.05	523	2.31
8	*s:Offer*	8,456	6.03	1,474	6.50	21	*d:Review*	2,816	2.01	783	3.45
9	*d:Rating*	7,711	5.50	1,726	7.61	..	..	..	..	..	..
10	*s:AggregateRating*	7,029	5.01	1,791	7.90	..	..	..	..	..	..
11	*s:Organization*	7,011	5.00	1,270	5.60	26	*d:Offer*	1,957	1.39	670	2.96
12	*d:Product*	6,770	4.82	1,156	5.10	31	*s:NewsArticle*	1,047	0.75	346	1.53
13	*s:WebPage*	6,678	4.76	2,112	9.32	41	*s:JobPosting*	552	0.39	154	0.68

Table 6. Most frequently used Microdata classes. From reasons of space *schema:* is cut short with *s:* and *data-vocabulary.org* with *d:* in this table.

Blog and Document Metadata: The class *schema:BlogPosting* is used by 25,235 websites making it the single most used Microdata class in our data set. The usage rate of this class on all websites containing Microdata is 17.98%, meaning that it is about three times higher than the usage rate on websites that belong to the Alexa 1 million list. The class *schema:Article* is used on 15,718 websites (13.35%). This class is a superclass of *schema:BlogPosting* and can be used to identify any kind of articles including newspaper and magazine articles. *Schema.org* also offers a specific class for annotating news articles. This class, *schema:NewsArticle*, was introduced in 2011 as a results of a collaboration between *schema.org* and the International Press Telecommunication Council, including companies like the New York Times, see [2]. This more specific class is used by 1,047 websites within our corpus, see Table 6. Using co-occurrence analysis, we found *schema:NewsArticle* to be used mostly in an isolated manner: Less than 1% of the websites that use *schema:NewsArticle* also use a second class (e.g. *schema:Person*).

Regarding the properties which are used together with *schema:NewsArticle*, we discovered that in around 79% of the cases the title property is filled and on 66% of the websites the *schema:articleBody* is used together with the class.

Navigational Information: The second most frequently used Microdata class is *dv:Breadcrumb* which is used by 21,729 websites. Breadcrumbs describe the navigational structure of a website. The Breadcrumb data is used by search engines to provide shortcuts to sub-pages within their search result. The class is used by 23.13% of the Alexa 1 Million websites using Microdata, but only by 15.49% of all websites using Microdata, meaning that popular websites tend to employ *datavoc:Breadcrumb* more frequently than others. A similar observation can be made for *schema:WebPage* which describes a web page and can also contain navigational information via its *schema:WebPage/breadcrumb* property.

Product Data: Reviewing all Microdata classes that are used by more than 100 different websites, we could identify four classes, i.e. *schema:Product*, *schema:Offer*, *dv:Product*, and *dv:Offer*, that describe products and product offers, see Table 6. The most frequently used product-related Microdata class is *schema:Product* which is used by 16,612 websites and thus shows a similar adoption level than the top product-related RDFa class, *og:"product"*, see Section 5. Websites using the product-related classes in the *dv* namespace employ on average 17.5 different properties, while sites using the product classes in the *schema.org* namespace on average only make use of 12 different properties. Table 7 lists the properties that are commonly used to describe *schema:Product*s as well as other product-related classes. *schema:Product/name*, *schema:Product/description*, *schema:Product/image*, and *schema:Product/offers* are the most frequently used properties which are used by at least 45.42% of the websites. The 26 other properties that are defined by *schema.org* for describing products are used by significantly less websites.

		PLDs Total	
	Property	#	%
1	*schema:Product/name*	14,342	86.34
2	*schema:Product/description*	10,297	61.99
3	*schema:Product/image*	8,093	48.72
4	*schema:Product/offers*	7,545	45.42
5	*schema:Offer/price*	6,894	41.50
6	*schema:AggregateRating/ratingValue*	3,990	24.02
7	*schema:PostalAddress/streetAddress*	3,723	22.41
8	*schema:PostalAddress/addressRegion*	3,502	21.08
9	*schema:PostalAddress/addressLocality*	3,074	18.50
10	*schema:LocalBusiness/address*	2,797	16.84

Table 7. Top properties that are used to describe *schema:Product*s as well as other product-related classes.

We further investigated which other classes are used together with *schema:Product* on the same website. The results are presented in Table 8 and reveal that only 43.31% of the websites use *schema:Product* together with *schema:Offer*, the Schema.org class

for representing offer details like *schema:Offer/price* or *schema:Offer/priceCurrency*. 25.93% of the websites provide *schema:AggregateRating*s for their products.

Class co-occurrence	# PLDs
{*schema:Product*}	16,612
{*schema:Product*, *schema:Offer*}	7,194
{*schema:Product*, *schema:AggregateRating*}	4,308
{*schema:Product*, *schema:Offer*, *schema:AggregateRating*}	3,226
{*schema:Product*, *dv:Product*}	2,810
{*schema:Product*, *schema:Offer*, *dv:Product*, *dv:Offer-aggregate*}	2,701

Table 8. Absolute PLD count for the 6 classes most frequently co-occurring with *schema:product*.

Ratings: The schema.org vocabulary offers two classes for representing rating information: *schema:Rating* for representing individual ratings and *schema:AggregateRating* for representing summaries of multiple ratings. Within our corpus, 7,000 websites provide aggregate ratings while only 1,532 websites markup the rating values of individual reviews. Aggregate ratings refer to *schema:Product* on around 1/3 of the websites, followed by *schema:LocalBusiness*es which are rated on 20% of the websites, and *schema:WebPage*s which are rated on around 10% of the websites. Sites using one of the rating classes provide in average 19 to 20 properties on their pages. Examining the rating scales, we found that most websites use a 0-to-5 scale with the values 5, 4 and 0 being used most frequently. *schema:Rating* refers to *schema:Product* on almost 50% of the 1.532 websites, followed by *schema:SoftwareApplication* (8%) and *schema:LocalBusiness* (7%).

Business Listings: The fifth most common Microdata class is *schema:LocalBusiness* which is used by 16,383 websites (11.68% of all websites containing Microdata). The class is used to describe a physical business like a shop or restaurant. 61.14% of the websites that use *schema:LocalBusiness* also provide a *schema:PostalAddress* for the business. The second most frequently co-occurring class is *schema:Product* (17.10%).

Job Postings: Resulting from a collaboration with the United States Office of Science and Technology Policy, schema.org started to provide vocabulary terms for describing job postings in the end of 2011, see [4]. We found 552 websites to use the *schema:JobPostings* class. Among the websites using *schema:JobPostings*, almost all websites (94.75%) also provide job titles (*schema:JobPosting/title*). About 50% of the websites make use of the properties *schema:JobPosting/jobLocation* and *schema:JobPosting/description*, and 40% give information about the hiring organizations using the property *schema:JobPosting/hiringOrganization*. Although *schema.org* defines the range of *schema:JobPosting/hiringOrganization* to be *schema:Organization*, over 60% of the websites use literals (like 'IBM' and 'eBay') instead of instances of the class organization to identify the hiring organization. Other more specific properties to describe *schema:JobPostings* such as *schema:JobPosting/skills* or *schema:JobPosting/benefits* are

rarely used. The property *schema:JobPosting/skills* is used only by 10% of all websites providing job postings and the property *schema:JobPosting/benefits* only by 2%.

7 Microformat Deployment

Microformats are used on approximately 1.7 million websites making them the most widely adopted markup format. 14.18% of the websites in the Alexa 1 Million list employ Microformats (see Table 2). Examples of websites from the Alexa Top 100 list that use Microformats are the online encyclopedia Wikipedia, which uses a large number of different Microformats, the Adobe website, and the Taobao marketplace, one of the most popular Chinese customer-to-customer online marketplaces.

Frequent Classes: Table 9 shows the most frequently used Microformat classes. The last two columns of the table contain the number and percentage of Microformat websites that are included in the Alexa top 1 million list. The table shows that hCard is by far the most widely used Microformat. Among others, hCard is used by the two micro-blogging platforms tumblr and twitter. The hCard type VCard is found in over 84% of all websites that use Microformats, followed by the hCard sub-classes Organization and Location. The second most widely deployed Microformat is hCalendar which is used by around 37 thousand websites. This format is among others used by the networking platform LinkedIn. The Microformats XFN and geo do not define classes and are thus not included into Table 9. Almost half a million websites use XFN while 48 thousand contain geo markup.

		PLDs Total		PLDs in Alexa	
	Class	#	%	#	%
1	*hCard:VCard*	1,511,467	84.03	87,758	83.79
2	*hCard:Organization*	195,493	10.87	10,430	9.96
3	*hCard:Location*	48,415	2.69	2,784	2.66
4	*hCalendar:vcalendar*	37,620	2.09	4,614	4.41
5	*hCalendar:Vevent*	36,349	2.02	4,400	4.20
6	*hReview:Review*	20,781	1.16	3,659	3.49
7	*hListing:Lister*	4,030	0.22	244	0.23
8	*hListing:Listing*	4,030	0.22	244	0.23
9	*hRecipe:Recipe*	3,281	0.18	1,068	1.02
10	*hListing:Item*	2,957	0.16	164	0.16
11	*hRecipe:Ingredient*	2,658	0.15	891	0.85
12	*hRecipe:Duration*	1,323	0.07	473	0.45
13	*hRecipe:Nutrition*	818	0.05	300	0.29
14	*species:species*	91	0.01	38	0.04
15	*species:Genus*	61	0.00	24	0.02
16	*species:Family*	60	0.00	24	0.02
17	*species:Kingdom*	59	0.00	24	0.02
18	*species:Order*	59	0.00	25	0.02

Table 9. Most frequently used Microformats classes.

Co-occurrence of Microformats on the same Website: 1.5 million websites use only a single Microformat (83% of all websites using Microformats). Almost 300 thousand (17%) websites use 2 formats, 9,428 (less than 1%) use 3 formats, 1,348 use 4 formats and 123 use 5 different formats. 30 websites use more than 5 different Microformats (for instance *blogspot.com*). Table 10 shows the most frequently co-occurring Microformats. It is noticeable that hCard is used together with most of the other Microformats. This fact is not really surprising as most other Microformat specifications rely on hCard for describing persons or organizations.

Microformats Co-occurrence	# PLDs
hcard, xfn	230,551
geo, hcard	35,341
hcalendar, hcard	10,508
hcard, hreview	7,858
hcalendar, hcard, xfn	2,104
geo, hcard, xfn	1,800
hcard, hlisting	1,742
geo, hcard, hreview	1,366

Table 10. Absolute PLD count of most frequently co-occurring Microformats used on the same PLD.

Property	# PLDs	% PLDs
1 *hCard:n*	1,511,467	100.00
2 *hCard:fn*	1,322,359	87.49
3 *hCard:url*	976,967	64.62
4 *hCard:photo*	413,613	27.36
5 *xfn:mePage*	239,240	15.83

Table 11. Absolute and relative PLD count of the most frequently used *hCard:VCard* properties.

HCard and XFN: The hCard root-class, *VCard* is used on over 1.5 million websites. The hCard class *Organization* is used most frequently together with *VCard* on 195,493 websites. Table 11 lists the Top 5 properties that are used to represent *VCard* information. We see that the provided descriptions are rather shallow and mostly only consist of a name and maybe a link to a person's homepage.

The second most frequently deployed Microformat is XFN, which is used by 490,286 websites within our data set. Using a co-occurrence analysis, we discovered that in almost 50% of all websites XFN relations (e.g. *xfn:mePage*) are used together with *VCard* classes. To analyze which websites or systems do support XFN, we extracted some of the pages containing relevant structured data. We found out that for instance Wordpress[16] automatically publishes XFN when users link to other blogs or friends' websites.

HCalendar: 37 thousand websites offer information using hCalendar. Out of these websites, 44% also use the Microformat hCard, in particular the class *VCard* in order to identify for instance event *attendee*s or *organizer*s.

HListing: HListing is a Microformat for annotating small-ads and classifieds. The format is used on 4,030 websites. From the websites employing hListing classes, around 80% also offer information about the price and over 70% do use the optional property item with the two properties *itemUrl* and *itemPhoto*. Overall we found almost 3 thousand websites to offer detailed information about a listing (*lister*, *item*, *price*, *itemUrl* and *itemPhoto*).

[16] `http://wordpress.com/`

HRecipe: The hRecipe Microformat is used to annotate cooking recipes on websites. We identified 3,278 websites offering structured data about recipes. Over 80% do list ingredients for their recipes and 20% of the sites offer additional information like durations and nutrition information. 40% of the websites use hCard together with hRecipe in order to include information about the authors of the recipes.

8 Related Work

In [9], Mika and Potter present an analysis of the deployment of RDFa, Microdata and Microformats based on a sample of the crawl of the Bing search engine (3.2 billion URLs, January 2012). The results of their study are mostly in line with our findings. For instance, they identified structured data on 4.7% of the examined websites while we found structured data on 5.64% of the websites. All our RDFa top classes listed in (Table 4) are also contained in their top 20 RDFa classes (without considering the Open Graph Protocol *og:* types, as Mika and Potter do not count them as RDFa classes). The findings also differ in some points, as both crawls obviously are only subsets of the whole Web and as the results are influenced by the crawling strategy employed by the two different crawlers that were used to gather the corpora. An example of diverting results is the number of websites that use *foaf:Image*: Mika and Potter report 30,903 websites for *foaf:Image*, compared to 44,644 websites according to our extraction. The analysis presented in this paper goes beyond the analysis presented by Mika and Potter, as we also analyze which properties are used to describe instances of popular classes as well as the co-occurrence of classes and thus also provide an indication about the richness and usefulness of the published data. A further difference between our work and the study by Mika and Potter is that their results are not verifiable as the Bing crawl is not publicly accessible. In contrast, the Common Crawl corpus, as well as our extracted data is available for download and can be used for further research.

In [8], Mika presents statistics about RDFa and Microformats distribution based on crawls from Yahoo!. The crawls date from 2008 to 2010 and are thus older than the corpus analyzed in this paper. The numbers given in [8] are not aggregated by website and thus depend highly on the crawling strategy of the Yahoo! crawler. Additional vocabulary-level statistics for the same Yahoo! crawl are provided by the W3C[17]. The statistics confirm our finding on the wide adoption of the Open Graph Protocol.

The commercial company BuiltWith[18] collects statistics about the deployment of RDFa and Microdata on 1 million popular PLDs. They report 166,000 websites to contain RDFa while we discovered 519,000. For Microdata, they found 295,000 websites while our data set only contains 140,000. As BuiltWith sells the lists of the websites containing structured data, verifying their results is expensive.

The Sindice search engine[19] collects data from the Web and allows the data to be searched using keyword queries and to be queried using SPARQL. Sindice only extracts data from the HTML pages of websites that provide site maps. In addition to data from HTML pages, Sindice also extracts data from WebAPIs and loads data sets from the

[17] http://www.w3.org/2010/02/rdfa/profile/data/yahoo/
[18] http://trends.builtwith.com
[19] http://sindice.com/

Linked Data Cloud. Sindice mixes this data with the HTML-extracted data in its index. Statistics[20] about the Sindice index are thus not directly comparable with the results presented in this paper. While we focus on wide coverage, Sindice focuses on deeper crawling. Consequently, the Sindice index covers less websites than Web Data Commons, especially for Microdata and Microformats. According to the Sindice statistics from March 30, 2013, the index contains RDFa from 420,409 websites, Microdata from 20,920 websites and Microformats from 295,262 websites (HCard).

We already presented the Web Data Commons project and a preliminary analysis of the extracted data as a short paper at the LDOW2012 workshop [7]. Compared to the LDOW2012 paper, the analysis presented in this paper is based on a larger web crawl (3 billion pages vs. 1.5 billion). The former paper did not present any class/property co-occurrence analysis and also did not aggregate the extracted data by PLD, meaning that the presented results are largely influenced by the crawling strategy of the Common Crawl.

9 Conclusion

Our study has shown that RDFa, Microdata, and Microformats have all three found considerable adoption on the Web and are being used by hundreds of thousands of websites. The adoption is also global, as we were able to identify considerable amounts of websites using the formats on all examined top-level-domains. Matching the websites that provide structured data with the Alexa list of popular websites revealed that nearly 50% of the top 10,000 websites embed structured data.

Concerning the topical domains of the published data, we found out that the dominant domains are: persons and organizations (represented using all three formats), blog- and CMS-related metadata (represented using RDFa and Microdata), navigational metadata (represented using RDFa and Microdata), product data (represented using all three formats), and event data (represented using a Microformat). Additional topical domains with smaller adoption include job postings (represented using Microdata) and recipes (represented using a Microformat). The topics of the data, as well as the formats and vocabularies used to represent the data, seem to be largely determined by the major consumers the data is targeted at: Google, Facebook, Yahoo!, and Bing. For instance, the examined RDFa data is dominated by the vocabulary promoted by Facebook, while the examined Microdata is dominated by the vocabularies promoted by Google, Yahoo!, and Bing via *schema.org*.

Concerning the structural richness of the published data, we found out that many websites only use a small set of rather generic properties to describe entities. For example, instances of the Open Graph Protocol class product are described using only the properties *title*, *url*, *site_name* and *description* in most cases. The same is true for instances of *schema:Product* for which 61.99% of the websites only provide a name and a description despite of *schema.org* defining 26 additional properties to describe products. This means that applications that for instance want to find out which websites offer a specific product need to employ additional information extraction techniques on

[20] `http://sindice.com/stats/`

these fields in order to gain a deeper understanding of their content (exact product type, product features), following the promise that a little semantics goes a long way.

All data that we have extracted from the Common Crawl as well as further, more detailed statistics about the adoption of the different formats are provided on the WebDataCommons.org website. By publishing the extracted data, we hope on the one hand to initialize further domain-specific studies by third parties. On the other hand, we hope to lay the foundation for enlarging the number of applications that consume structured data from the Web, as the URLs of the webpages that we identified to contain a specific type of data can be used as seeds for topic-specific deeper crawls.

Acknowledgements

We would like to thank the Common Crawl foundation for publishing recent Web crawls as well as the Any23 team for their great parsing framework. This work has been supported by the LOD2 and PlanetData research projects funded by the European Community's Seventh Framework Programme. Johanna Völker is financed by a Margarete-von-Wrangell scholarship of the European Social Fund (ESF) and the Ministry of Science, Research and Arts Baden-Württemberg.

References

1. B. Adida and M. Birbeck. RDFa primer - bridging the human and data webs - W3C recommendation. http://www.w3.org/TR/xhtml-rdfa-primer/, 2008.
2. K. Goel. Extended schema.org news support. http://blog.schema.org/2011/09/extended-schemaorg-news-support.html, 2011.
3. K. Goel, R. V. Guha, and O. Hansson. Introducing rich snippets. http://googlewebmastercentral.blogspot.de/2009/05/introducing-rich-snippets.html, 2009.
4. R. V. Guha. Schema.org support for job postings. http://blog.schema.org/2011/11/schemaorg-support-for-job-postings.html, 2011.
5. K. Haas, P. Mika, P. Tarjan, and R. Blanco. Enhanced results for web search. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 725–734, New York, NY, USA, 2011. ACM.
6. I. Hickson. HTML Microdata. http://www.w3.org/TR/microdata/, 2011. Working Draft.
7. H. Mühleisen and C. Bizer. Web data commons – extracting structured data from two large web corpora. In *LDOW 2012: Linked Data on the Web*, CEUR Workshop Proceedings, Vol. 937. CEUR-ws.org, 2012.
8. P. Mika. Microformats and RDFa deployment across the Web . http://tripletalk.wordpress.com/2011/01/25/rdfa-deployment-across-the-web/, 2011.
9. P. Mika and T. Potter. Metadata statistics for a large web corpus. In *LDOW 2012: Linked Data on the Web*, CEUR Workshop Proceedings, Vol. 937. CEUR-ws.org, 2012.