The WebDataCommons Microdata, RDFa and Microformat Dataset Series

Robert Meusel, Petar Petrovski, and Christian Bizer

University of Mannheim, Germany
Research Group Data and Web Science
{robert,petar,chris}@informatik.uni-mannheim.de

Abstract. In order to support web applications to understand the content of HTML pages an increasing number of websites have started to annotate structured data within their pages using markup formats such as Microdata, RDFa, Microformats. The annotations are used by Google, Yahoo!, Yandex, Bing and Facebook to enrich search results and to display entity descriptions within their applications. In this paper, we present a series of publicly accessible Microdata, RDFa, Microformats datasets that we have extracted from three large web corpora dating from 2010, 2012 and 2013. Altogether, the datasets consist of almost 30 billion RDF quads. The most recent of the datasets contains amongst other data over 211 million product descriptions, 54 million reviews and 125 million postal addresses originating from thousands of websites. The availability of the datasets lays the foundation for further research on integrating and cleansing the data as well as for exploring its utility within different application contexts. As the dataset series covers four years, it can also be used to analyze the evolution of the adoption of the markup formats.

Keywords: Microdata, RDFa, Microformats, Dataset, Web Science.

1 Introduction

A large number of websites have started to use markup standards to annotate information about products, reviews, blog posts, people, organizations, events, and cooking recipes within their HTML pages. The most prevalent of these standards are Microformats,[1] 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 [7], 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!, Yandex, and Bing, which use the data to enrich search results and to display entity descriptions within their applications [6,3]. Since 2011, those four search engine companies have been collaborating on the

[1] http://microformats.org/

P. Mika et al. (Eds.) ISWC 2014, Part I, LNCS 8796, pp. 277–292, 2014.

Schema.org initiative,[2] which offers a single vocabulary for describing entities that is understood by applications from all four companies [5]. So far, only the big search engine companies had access to large quantities of Microdata, RDFa, and Microformats data as they were the only ones possessing large web crawls. However, the situation has changed with the advent of the *Common Crawl Foundation*.[3] Common Crawl is a non-profit foundation that crawls the Web and regularly publishes the resulting web corpora for public usage.

We have extracted all Microdata, RDFa, and Microformats data from the Common Crawl corpora gathered in 2010, 2012 and 2013 and provide the extracted data for public download. Table 1 gives an overview of the Common Crawl corpora as well as the overall quantity of the extracted data. The second and third column show the number of HTML pages and pay-level domains (PLDs) covered by the different crawls. The forth and fifth column contain the percentages of all pages and PLDs that use at least one of the three markup formats. Column six shows the overall number of RDF quads that we have extracted from each corpus, while column seven contains the compressed size of the resulting datasets. The 2013 Common Crawl corpus, for instance, consists of 2.2 billion HTML pages originating from over 12 million PLDs. 26.33% of these pages and 13.87% of the PLDs use at least one markup format, resulting in an extracted dataset containing 17 billion RDF quads.

Table 1. Overview of the Common Crawl corpora and the overall quantity of the extracted data

	Crawl Size		Extracted Data			
Dataset	# HTML Pages	# PLDs	% HTML Pages	% PLDs	# RDF Quads	Compressed Size
2010	2 565 741 671	-	5.76%	-	5 193 767 058	332 GB
2012	3 005 629 093	40 600 000	12.29%	5.63%	7 350 953 995	101 GB
2013	2 224 829 946	12 831 509	26.33%	13.87%	17 241 313 916	40 GB

This paper is structured as follows: first, we give an overview of the Common Crawl initiative and the web corpora that it provides to the public. Afterwards, we explain the methodology that was used to extract the data from the corpora and describe the data format that we use to offer the data for public download. In order to give an impression of the content of the extracted data, we discuss the distribution of the different markup formats within the 2013 dataset in Section 5. Afterwards, we analyze the topical domains as well as the richness of the annotations in Section 6 for RDFa, Section 7 for Microdata, and Section 8 for Microformats. In [2], we have presented a similar analysis of the 2012 dataset. In order to illustrate the evolution of the adoption of the different formats, we compare our findings from the 2012 and 2013 datasets wherever this reveals interesting trends. Section 9 discusses related work, while Section 10 concludes the paper by discussing the challenges that need to be addressed for using the data within applications.

[2] `http://schema.org`
[3] `http://commoncrawl.org`

2 The Common Crawl

Our dataset series was extracted from three web corpora published by the Common Crawl Foundation. The first corpus contains pages that have been crawled between 2009 and 2010. The second corpus was gathered in the first half of 2012. The crawler that was used to gather both corpora employed a breath-first selection strategy and was seeded with a large number of URLs from former crawls. The seed URLs were ordered according to their PageRank. Since the end of 2012 the Common Crawl Foundation releases two crawls per year. Each crawl consists of around two billion pages. For the recent crawls the foundation uses seed lists provided by the search engine company blekko.[4] The new seed lists should improve the quality of the crawl by avoiding "webspam, porn and the influence of excessive SEO" [8]. In addition to using an external seed list, the Common Crawl Foundation has also shifted their crawling infrastructure to a modified version of Apache Nutch to gather the pages contained in the seed list instead of using their own crawling framework.[5] All Common Crawl corpora are provided as (W)ARC files[6] and are available as free download from Amazon S3.[7]

3 Methodology

In order to extract RDFa, Microdata, and Microformats data from the corpora, we developed a parsing framework which can be executed on Amazon EC2[8] and supports parallel processing of multiple (W)ARC files. The framework relies on the *Anything To Triples* parser library (Any23)[9] to extract Microdata, RDFa, and Microformats data from the corpora. For processing the Common Crawl corpora on Amazon EC2 we used 100 AWS EC2 *c1.xlarge* machines. Altogether, extracting the HTML-embedded data from the 2013 corpus required a total machine rental fee of US$ 263.06 using Amazon spot instances.[10]

We used Apache Pig[11] running on Amazon Elastic MapReduce to calculate most of the statistics presented in this paper as well as to generate the vector representation used for the co-occurrence analysis.[12] As the three crawls cover different HTML pages and as the number of crawled pages per PLD differs

[4] `http://blekko.com/`

[5] The code which was used for the crawl can be downloaded at `https://github.com/Aloisius` and the original distribution of Nutch at `https://nutch.apache.org/`

[6] The WARC file format is proposed by the Internet Archive foundation as successor to the ARC file format – `http://archive-access.sourceforge.net/warc/`.

[7] `http://aws.amazon.com/datasets/41740`

[8] `http://aws.amazon.com/de/ec2/`

[9] `http://any23.apache.org/`

[10] Additional information about the extraction framework can be found at `http://webdatacommons.org/framework`

[11] `http://pig.apache.org/`

[12] All used scripts can also be downloaded from the websites of the Web Data Commons project.

widely, we aggregate the data by PLD, especially for analyzing the deployment of the different markup languages and comparing the deployment between the different datasets. To determine the PLD of each page, we use the Public Suffix List.[13] Hence, a PLD not always equals a second-level domain, but country-specific domains such as "co.uk" or mass hosting domains like *blogspot.com* are considered as top-level domains in our analysis.

4 Dataset Format and Download

The extracted data is represented as RDF quads (encoded as N-Quads[14]), with the forth element being used to represent the provenance of each triple. This means in addition to subject, predicate, and object, each quad includes the URL of the HTML page from which it was extracted. The extracted data is provided for download in the various sub-datasets. Each sub-dataset includes the information extracted for one markup language from one crawl, e.g. all quads representing information embedded in web pages from the 2013 crawl using Microdata form a sub-dataset. All datasets are provided for public download on the Web Data Commons website.[15] In addition to the datasets, the website also provides detailed background data for the analysis presented in this paper, such as the lists of all websites using specific formats or vocabulary terms.

5 Distribution by Format

Table 2 gives an overview of the distribution of the different markup formats within the 2013 dataset. For each format, the table contains the number of PLDs and the number of URLs using the format. For Microformats, the numbers are reported separately for each sub-format. Column 5 and 6 contain the number of quads and the compressed file size of the extracted datasets. The largest number of quads, namely 8.7 billion, were generated from Microdata annotations, followed by the Microformat hcard with 4.9 billion and RDFa with over 2.6 billion quads. Regarding the number of websites annotating information using the different markup languages, we find 995 thousand websites using hcard, followed by 471 thousand using RDFa and 463 thousand using Microdata.

In order to give an impression about the number of entities that are described in the data as well as the richness of the entity descriptions, we group all quads that have the same subject URI into a *record*. Column four of Table 2 contains the overall number of records contained in each dataset. We see, for instance, that the Microdata dataset describes 1.9 billion entities. Each entity description (record) consists of an average of 4.48 quads.

[13] `http://publicsuffix.org/list/`
[14] `http://sw.deri.org/2008/07/n-quads/`
[15] `http://webdatacommons.org/structureddata/`

Table 2. Number of websites (PLDs) and webpages (URLs) containing RDFa, Microdata, and Microformats annotations, as well as number of records and quads within the 2013 dataset

	# PLDs	# URLs	# Records	# Quads	File Size
RDFa	471 406	296 005 115	436 100 210	2 636 964 693	66 GB
Microdata	463 539	276 348 609	1 964 777 851	8 795 074 538	189 GB
Microformats (geo)	23 044	14 436 467	56 611 312	222 780 517	4 GB
Microformats (hcalendar)	20 981	3 683 002	41 683 362	212 675 776	2 GB
Microformats (hcard)	995 258	113 402 968	1 643 288 889	4 884 918 863	60 GB
Microformats (hlisting)	2 854	528 387	19 204 882	65 494 465	890 MB
Microformats (hrecipe)	3 539	814 793	7 094 914	34 062 142	890 MB
Microformats (hresume)	262	52 675	81 924	231 573	4 MB
Microformats (hreview)	12 880	3 504 643	33 027 023	145 692 102	4 GB
Microformats (species)	109	22 419	121 200	373 033	6 MB
Microformats (xfn)	195 663	18 467 168	62 571 191	243 046 214	2 GB

6 RDFa Data

The 2013 RDFa dataset includes data from over 471 thousand websites, which are 26% of all websites containing structured data in the crawl. The largest amount of RDF statements was extracted from *tripadvisor.com* with 78 million quads, followed by *yahoo.com* with over 28 million quads and *hotels.com* with more than 17 million quads.

Class/Property Frequency Distribution: The corpus contains over 646 thousand different classes and over 27 thousand different RDFa properties. Figure 1(a) shows the class and property distribution using a log-scale for the y-axis, which reports the number of websites making use of a class or property. The x-axis draws the classes and properties ordered descending by the number of websites using them. Similar to our observations for the 2012 dataset [2], both distributions are long-tailed and only a small number of classes and properties are used by a large number of websites. Altogether, we find 949 classes and 2 069 properties that are used by at least two different websites. The majority of the terms are only used by a single website. Manually inspecting some of these terms reveals a large number of typos in spelling terms from more widely used vocabularies. On the other hand, there exists also a large number of proprietary vocabularies which are used only by a single website.

Frequent Classes: Table 3 lists the most frequently used RDFa classes ordered by the number of websites deploying them. The table also includes the total number of records of each class included in the 2013 dataset. For comparison, we also state the total as well as the percental number of websites deploying the classes in 2012.[16] Table 3 shows that the *Facebook* ecosystem has a strong presence in the most frequently used classes, i.e. nine out of 30 classes belong to the Open Graph Protocol (OGP). Although the total number of websites using

[16] The namespaces of the classes are abbreviated with the corresponding prefix from the `http://prefix.cc/list`. Classes with an *og*-namespace prefix belong to the OGP and are within the HTML pages not maintained with a namespace, but as literals instead.

the classes *og:"article"* and *og:"website"* is smaller in the 2013 dataset than in the 2012 dataset, the percental usage is higher. This is due to the smaller number of PLDs covered in the 2013 crawl (see Table 1). Looking at the total number of records of each class (column 3 in Table 3), we see that the dataset contains 13 million *og:"product"* records, 15 million *gd:Organization* records, as well as 22 million *sioc:UserAccount* records.

Table 3. Most frequently used RDFa classes within the 2013 dataset sorted by the number of websites (PLDs) using the class, including the total number of records in 2013 as well as the number of websites using the class in 2012

		2013			2012		
		Records	PLDs		Records	PLDs	
	Class	# (in k)	#	%	# (in k)	#	%
1	og:"article"	82 882 535	167 544	40.14	35 438 354	183 046	35.24
2	og:"website"	24 951 292	71 590	17.15	9 197 072	56 573	10.89
3	foaf:Image	143 179 835	46 505	11.14	12 618 426	44 644	8.60
4	foaf:Document	31 601 886	45 542	10.91	3 709 728	49 252	9.48
5	gd:Breadcrumb	53 156 451	39 561	9.48	52 521 380	9 054	1.74
6	og:"blog"	6 364 724	29 629	7.10	2 365 037	58 971	11.35
7	sioc:Item	30 863 230	29 521	7.07	3 325 019	33 141	6.38
8	og:"product"	13 199 034	13 813	3.31	7 517 484	19 107	3.68
9	sioc:UserAccount	22 195 639	12 632	3.03	2 067 204	19 331	3.72
10	skos:Concept	24 011 250	11 873	2.84	5 197 930	13 477	2.59
11	gd:Review-aggregate	16 626 171	5 266	1.26	7 419 398	6 236	1.20
12	sioc:Post	26 571 378	4 958	1.19	1 079 844	6 994	1.35
13	gd:Rating	979 322	3 603	0.86	1 567 226	4 139	0.80
14	og:"company"	1 834 688	3 105	0.74	2 483 995	6 758	1.30
15	sioctypes:BlogPost	653 322	2 703	0.65	159 553	3 936	0.76
16	sioctypes:Comment	25 831 008	2 639	0.63	903 696	3 339	0.64
17	vcard:Address	55 425	2 225	0.53	746 673	3 167	0.61
18	gr:Offering	498 333	2 199	0.53	371 864	1 342	0.26
19	gr:BusinessEnttiy	394 556	2 155	0.52	119 394	3 155	0.61
20	og:"activity"	1 049 085	2 037	0.49	913 007	3 303	0.64
21	gr:UnitPriceSpecification	429 409	1 681	0.40	450 220	1 562	0.30
22	gr:SomeItems	235 785	1 429	0.34	148 689	670	0.13
23	og:"profile"	940 016	1 276	0.31	573 848	394	0.08
24	gd:Organization	15 693 269	1 232	0.30	7 324 570	2 502	0.48
25	gd:Review	1 415 844	1 221	0.29	1 085	1 321	0.25
26	og:"band"	106 524	1 168	0.28	468 385	1 988	0.38
27	og:"game"	679 546	1 123	0.27	936 482	1 336	0.26
28	gr:TypeAndQuantityNode	187 865	1 121	0.27	122 137	530	0.10
29	gr:QuantitativeValue	192 560	1 032	0.25	282 325	1 077	0.21
30	foaf:Person	1 338 823	851	0.20	128 475	1 209	0.23

Facebook Data: In the following we will have a brief look at the OGP data and state properties included in the dataset for the OGP classes. The OGP is developed and promoted by *Facebook* in order to enable the integration of external content into the social networking platform. In contrast to other RDFa vocabularies, OGP allows the usage of literals instead of URIs to identify classes. Table 4 shows the properties that are most frequently used together with the top five OGP classes. Similar to our findings for the 2012 dataset [2], the top 15 most frequently used properties are rather generic, whereas there is a small shift in the usage of namespaces as the *ogm* namespace is used more frequently.

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

	og:"article"		og:"website"		og:"blog"		og:"product"		og:"company"	
Property	#	%	#	%	#	%	#	%	#	%
ogo:type	116 898	69.77	32 034	44.75	15 534	52.43	9 909	71.74	1 096	35.30
ogo:title	115 867	69.16	31 737	44.33	15 024	50.71	9 845	71.27	985	31.72
ogo:url	115 508	68.94	31 416	43.88	15 224	51.38	9 662	69.95	965	31.08
ogo:site_name	109 888	65.59	27 088	37.84	15 365	51.86	9 709	70.29	963	31.01
ogo:image	92 874	55.43	23 567	32.92	9 716	32.79	9 793	70.90	921	29.66
ogo:description	80 209	47.87	25 258	35.28	10 931	36.89	9 157	66.29	729	23.49
ogm:type	49 631	29.62	39 347	54.96	14 122	47.66	3 785	27.40	2 017	64.96
ogm:title	49 152	29.34	38 292	53.49	13 982	47.19	3 697	26.76	1 978	63.70
ogm:url	48 769	29.11	37 784	52.78	13 931	47.02	3 578	25.90	1 904	61.32
ogm:site_name	46 865	27.97	31 234	43.63	13 880	46.85	3 241	23.46	1 847	59.49
ogm:description	42 068	25.11	28 499	39.81	11 501	38.82	3 020	21.86	1 667	53.70
ogm:image	36 923	22.04	26 300	36.74	9 983	33.69	3 540	25.63	1 863	60.00
fb_2008:fbmlapp_id	27 865	16.63	11 550	16.13	10 769	36.35	2 275	16.47	812	26.16
ogo:locale	24 200	14.44	14 809	20.69	4 731	15.97	126	0.91	103	3.32
fb_2008:fbmladmins	22 773	13.59	11 097	15.50	10 076	34.01	2 796	20.24	1 351	43.52

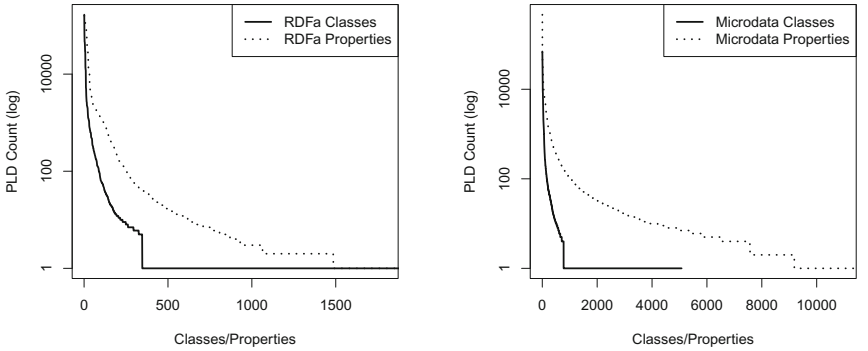
7 Microdata

The 2013 Microdata dataset contains data from over 463 thousand different websites, which are 26% of all websites containing structured data. Compared to the 6.1% of all websites using Microdata in 2012 [2], the adoption has grown by more than factor four in just one year. The largest amounts of Microdata statements were extracted from *citysearch.com* with 797 million quads, *ebay.com* with 153 million quads and *hp.com* with 65 million quads.

Class/Property Frequency Distribution: The dataset contains over 15 thousand different classes and over 170 thousand different properties that are used by Microdata annotations. Figure 1(b) shows the class and property distribution using a log-scale in the same manner as Figure 1(a). Altogether, the Microdata dataset contains 1 200 classes and 12 506 properties that are used by at least two different websites. Similar to the observations made for the RDFa deployment, classes and properties in the long tail include large numbers of typos as well as website-specific terms.

Frequent Classes: Table 5 shows the most frequently used Microdata classes ordered by the number of PLDs deploying them. The second column shows the absolute number of records of each class. The most popular classes belong to the topical domains product data (Product, Offer, Review, Rating), blogs (Article, Blog, BlogPosting), navigational information (Breadcrumb), people (Person), organizations (LocalBusiness, Organization) and addresses (PostalAddress, Address). Due to the growing adoption of Microdata, we discuss some of the major topical domains of the data in more detail in the following.

Postal Addresses: The dataset contains 124 million *schema:PostalAddress* records originating from over 52 thousand websites. On average each address is described by 3.96 property values. Table 6(a) shows that more than 90% of the



(a) RDFa (b) Microdata

Fig. 1. Class and Property distribution by PLD count within the 2013 dataset

records contain the properties *schema:addressLocality* and *schema:addressRegion*. Table 7(a) shows the top ten websites ordered by number of address records that we have extracted from the sites.

Local Business: The dataset contains over 76 million records of type *schema: LocalBusiness* coming from 35 403 websites. On average *schema:LocalBusiness* records contain 5.22 properties. As shown in Table 6(b), over 80% of all records contain four out of the top five properties. This means, that for a large proportion of records we can expect information about the address of the business, the name, as well as the URL. When comparing the websites using the *schema:LocalBusiness* class (cf. Table 7(b)) with the ones using the class *schema: PostalAddress* we see *citysearch.com* at the first position in both lists. The website is a local business search engine, providing information about companies within different cities. A remarkable observation for local businesses is that more than 6% of the records contain several values for the property "name".

Product Data: The 2013 dataset contains 202 million product records originating from almost 71 thousand different websites. This makes product data the second largest topical domain in the dataset. Table 7(c) shows the top ten PLDs offering product data ordered by the number of records. Product descriptions are markup with two different classes: *schema:Product* (80%) and *dv:Product* (20%). On average each product is described by 4.56 properties. Table 6(c) shows that the properties "name", "offers", and "image" are provided for almost 50% of all product records. Only around 17% of the records contain a "description" property. Only 15% of all records use of the property "productId" which might help to identify product records from different websites that refer to the same product. Petrovski *et al.* [12] have examined the content of product name properties of electronic products. Their analysis shows that there is quite some variation in the names that are used by different websites to refer to the same product and that many e-shops include different product features for marketing reasons

Table 5. Most frequently used Microdata classes within the 2013 dataset, sorted by the number of websites using the class, including the total number of records in 2013 as well as the number of websites using the class in 2012

		2013			2012		
		Records	PLDs		Records	PLDs	
	Class	# (in k)	#	%	# (in k)	#	%
1	schema:WebPage	33 806 314	69 712	15.04	5 927 825	6 678	4.76
2	schema:Article	53 456 896	65 930	14.22	5 012 240	15 718	11.20
3	schema:Blog	2 281 401	64 709	13.96	1 421 909	2 084	1.49
4	schema:Product	178 334 394	56 388	12.16	19 386 194	16 612	11.84
5	schema:PostalAddress	125 780 525	52 446	11.31	9 513 985	19 592	13.96
6	dv:Breadcrumb	223 814 124	44 187	9.53	75 537 788	21 729	15.49
7	schema:AggregateRating	47 467 552	36 823	7.94	4 446 934	7 029	5.01
8	schema:Offer	154 407 699	35 635	7.69	13 725 226	8 456	6.03
9	schema:LocalBusiness	76 317 387	35 264	7.61	7 467 891	16 383	11.68
10	schema:BlogPosting	5 505 020	32 056	6.92	12 143 573	25 235	17.98
11	schema:Organization	91 321 833	24 255	5.23	3 060 174	7 011	5.00
12	schema:Person	143 648 178	21 107	4.55	5 912 833	5 237	3.73
13	schema:ImageObject	32 712 837	16 084	3.47	5 404	283	0.20
14	dv:Product	19 990 466	13 844	2.99	6 235 638	6 770	4.82
15	schema:Review	35 213 270	13 137	2.83	3 114 006	2 585	1.84
16	dv:Review-aggregate	5 462 245	13 075	2.82	2 994 221	8 517	6.07
17	dv:Organization	4 951 153	9 582	2.07	2 311 548	5 853	4.17
18	dv:Offer	7 722 086	9 298	2.01	4 201 002	1 957	1.39
19	dv:Address	1 629 193	8 866	1.91	1 277 451	5 559	3.96
20	dv:Rating	5 878 816	8 360	1.80	2 063 366	1 532	1.09
21	schema:Event	10 551 937	8 258	1.78	1 018 398	4 102	2.92
22	schema:Place	38 519 652	7 653	1.65	1 819 200	4 131	2.94
23	dv:Review	1 868 702	6 432	1.39	1 019 152	2 816	2.01
24	schema:Recipe	1 523 363	6 019	1.30	379 433	718	0.51
25	schema:GeoCoordinates	72 961 757	5 888	1.27	1 045 302	4 677	3.33
26	schema:ProfilePage	116 065	4 833	1.04	86 572	30	0.02
27	schema:AutoDealer	49 706	4 563	0.98	31 615	280	0.20
28	schema:VideoObject	7 124 628	4 530	0.98	31 452 643	764	0.54
29	dv:Person	23 386 913	3 993	0.86	2 609 898	5 237	3.73
30	schema:Thing	1 214 435	3 724	0.80	141 641	587	0.42

into the product names. Both findings illustrate the difficulties that an application will need to face that tries to build an integrated product catalog based on Microdata product records. Petrovski *et al.* approach this problem by first extracting product features from the product names and descriptions and then using these features for identity resolution, reaching an F1-measure of 82% [12].

Job Postings: As a result of 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 [4]. Our dataset contains 21 million records of class *schema:JobPosting* originating from over two thousand websites. *schema:JobPosting* records contain, on average, 5.93 properties and the class *schema:JobPosting* thus belongs to the classes with the highest average number of properties used. Table 6(d) shows the most frequent properties of *schema:JobPosting* records. 1% of the records contain more than one "name" property value. Table 7(d) shows the top ten PLDs by record count providing data for job postings.[17]

[17] A complete list of websites that embed Microdata can be found at `http://www.webdatacommons.org/structureddata/2013-11/stats/stats.html#html-microdata`

Table 6. Most frequently used properties for selected classes. For space reasons, the *schema*-namespace prefix is shortened to *s* and class names are shortened according the respective heading.

(a) PostalAdress (PA) Records

Property	Records # (in k)	%
s:PA/addressLocality	122 008	98.07
s:PA/addressRegion	114 072	91.69
s:PA/streetAddress	81 719	65.69
s:PA/postalCode	25 447	20.45
s:PA/addressCountry	11 010	8.85
s:PA/telephone	2 790	2.21
s:PA/url	1 422	1.13
s:PA/AddressLocality	1 262	1.00
s:PA/AddressRegion	1 248	0.99
s:PA/name	615	0.49

(b) LocalBusiness (LB) Records

Property	Records # (in k)	%
s:LB/name	80 832	106.13
s:LB/address	70 427	92.47
s:LB/url	64 139	84.21
s:LB/geo	63 450	83.31
s:LB/telephone	9 165	12.03
s:LB/description	8 310	10.89
s:LB/image	8 115	10.63
s:LB/aggregateRating	4 320	5.66
s:LB/review	3 807	4.99
s:LB/openingHours	1 957	2.56

(c) Product (P) Records

Property	Records # (in k)	%
s:P/name	115 326	57.07
s:P/offers	112 826	55.83
s:P/image	96 193	47.60
s:P/url	59 848	29.62
s:P/description	34 334	16.99
s:P/productID	30 820	15.11
s:P/aggregateRating	24 832	12.17
s:P/image	24 082	11.81
s:P/brand	23 077	11.31
s:P/sku	14 637	7.18

(d) JobPosting (JP) Records

Property	Records # (in k)	%
s:JP/title	21 548	101.77
s:JP/hiringOrganization	20 539	97.01
s:JP/jobLocation	19 101	90.22
s:JP/description	14 877	70.27
s:JP/url	8 633	40.77
s:JP/name	8 283	39.12
s:JP/datePosted	5 578	26.35
s:JP/image	2 782	13.14
s:JP/skills	1 298	6.13
s:JP/address	606	2.86

7.1 New Microdata Adopters

In the following, we will analyze the websites that newly adopted Microdata in 2013. We use the list of websites extracted by Meusel *et al.* [9] from the 2012 crawl and calculate the overlap with the crawled websites in 2013. We then identify every website which is included in the 2012 and 2013 crawl and has adopted RDFa, Microdata, or Microformats in 2013 but did not adopt it in 2012. This results in a list of 490 778 websites out of which 169 134 make use of Microdata.

Table 8 gives an overview of the classes that are used by at least 1% of new adopters. Again, classes of the *Schema.org* vocabulary dominate, however despite its deprecation in 2011 the *data-vocabulary* vocabulary is still being used by the new adopters in 2013. Similar to the overall distribution of Microdata classes, websites newly adopting Microdata cover a broad range of different topics with a slight focus on product related data.

As an example, we calculated a co-occurrence matrix for classes and properties on websites newly adopting *schema:Product* and compare the co-occurring properties with the analysis of all *schema:Product* websites from the 2013 and 2012 datasets. Table 9 shows the top 20 most co-occurring properties on websites newly adopting Microdata. The table also shows in column six and eight the difference between the new adopters and the complete datasets from 2013 and 2012. Product records appearing on websites newly adopting Microdata are more likely described by the top six properties than in the overall dataset of

Table 7. Top ten PLDs ordered by number of Microdata records

(a) PostalAdress Records

Website	Records # (in k)	%
citysearch.com	61 623	49.53
peoplefinders.com	19 089	15.34
stubhub.com	4 921	3.96
seatgeek.com	4 205	3.38
viagogo.com	2 760	2.22
apartmentguide.com	2 299	1.85
monster.com	2 257	1.81
avvo.com	1 534	1.23
zillow.com	1 453	1.17
radaris.com	1 248	1.00

(b) LocalBusiness Records

Website	Records # (in k)	%
citysearch.com	64 297	84.42
yell.com	3 429	4.50
bbb.org	857	1.13
partypop.com	682	0.90
justia.com	343	0.45
vcahospitals.com	281	0.37
leisurepro.com	218	0.29
travelpod.com	215	0.28
vacationroost.com	196	0.26
nakedapartments.com	183	0.24

(c) Product Records

Website	Records # (in k)	%
ebay.com	18 362	9.09
fotolia.com	16 319	8.08
aliexpress.com	9 747	4.82
ebay.co.uk	8 600	4.26
competitivecyclist.com	5 549	2.75
swatch.com	5 199	2.57
ebay.ca	5 141	2.54
crateandbarrel.com	4 303	2.13
hp.com	4 018	1.99
bentgate.com	3 776	1.87

(d) JobPosting Records

Website	Records # (in k)	%
snagajob.com	5 899	27.86
indeed.com	4 176	19.72
startuphire.com	2 704	12.77
monster.com	2 418	11.42
simplyhired.com	1 847	8.73
glassdoor.com	1 492	7.05
itjobswatch.co.uk	522	2.47
spherion.com	109	0.52
glassdoor.ca	91	0.43
glassdoor.com.au	91	0.43

2013 and 2012. Further, this subset includes less rating information, but the records are more likely to contain information about the *manufacturer* and the *itemConditions*.

8 Microformats Data

Microformats are used on approximately 1.1 million websites within the 2013 crawl. This makes Microformats the most widely adapted markup format being used by over 62.7% of all sites using any markup languages.

Frequent Classes: Table 10 gives an overview of the most frequently used Microformats classes. The third column shows the absolute number of records of a certain class in the 2013 dataset. Column four shows the absolute number of PLDs from which the records originate. The last two columns show the percentage of PLDs making use of a certain Microformats classes in the 2013 and 2012 datasets. The most popular Microformat class is *hcard:VCard*. The dataset includes over 787 million records of this class originating from almost one million different sites. The second most frequent used class is *hCard:Organization*. The 2013 dataset contains over 126 million records of this class. Both classes belong to the *hCard* vocabulary. The second most frequently used Microformats vocabulary is *geo* with 75 million records of type *geo:Location* spread over 23 thousand sites. Besides the over 37 million *hCalendar:Vevents* records and 19 million *hReview:Review* records, the dataset also offers over one million recipes originating

Table 8. Microdata classes used by at least 1% of websites which newly annotate data using Microdata in 2013, ordered by the number of websites using them

	Class	PLDs #	PLDs %		Class	PLDs #	PLDs %
1	s:Product	28 198	16.67	15	dv:Offer	4 512	2.67
2	s:WebPage	27 672	16.36	16	s:Review	4 498	2.66
3	s:Article	23 908	14.14	17	http:/schema.orgStore	4 213	2.49
4	s:PostalAddress	22 731	13.44	18	dv:Organization	4 086	2.42
5	s:Offer	19 185	11.34	19	s:Event	3 969	2.35
6	dv:Breadcrumb	16 972	10.03	20	dv:Address	3 596	2.13
7	s:LocalBusiness	14 515	8.58	21	s:Place	3 417	2.02
8	s:AggregateRating	14 140	8.36	22	dv:Rating	2 770	1.64
9	s:Organization	11 123	6.58	23	s:ImageObject	2 690	1.59
10	s:Blog	9 780	5.78	24	s:Rating	2 503	1.48
11	s:Person	7 350	4.35	25	s:GeoCoordinates	2 387	1.41
12	s:BlogPosting	7 083	4.19	26	s:VideoObject	1 865	1.10
13	dv:Product	6 548	3.87	27	dv:Review	1 685	1.00
14	dv:Review-aggregate	4 782	2.83				

from 3 530 different sites. The top PLDs from which the data originates are *epicurious.com*, *grouprecipes.com* and *chefkoch.de*. Comparing the percentage of PLDs using Microformats annotations between the 2012 and 2013 datasets, the deployment of Microformats does not grow significantly but appears stable.

9 Related Work

In this section we review other public Microdata, RDFa, and Microformats datasets and refer to related work analyzing the deployment of these standards.

The only other public large-scale source of Microdata, RDFa, and Microformats data – that we are aware of – is the Sindice search engine.[18] Sindice collects data from the Web and allows the data to be searched using keyword as well as SPARQL queries. The Sindice index includes not only data gathered from HTML pages but also data extracted from WebAPIs as well as data from the Linked Data Cloud. The data is mixed by Sindice within their index which makes it difficult to get a pure HTML-extracted dataset. Also note that Sindice only crawls HTML pages from websites that offer a site map. According to the latest Sindice statistics from September 2013, their corpus contains 3.36 million different classes for which they could find at least six records within their data sources.[19] The index includes around 700 million records of class *hCard:VCard*, 68 million records of class *hCard:Organization*, 28 million records of class *og:article* and over 10 million records of class *schema:Product*. Unfortunately, according to recent Sindice blog posts, there are no plans to keep the SPARQL endpoint alive as well as to update their large datasets.[20] As Sindice is restricted to websites offering sitemaps, it does not cover as many websites as our datasets. On the other hand, Sindice covers websites in a more complete

[18] `http://sindice.com`

[19] `http://sindice.com/stats/direct/basic-class-stats`

[20] `https://groups.google.com/forum/#!topic/sindice-dev/ASzK-hKzNFA`

Table 9. Top properties that are used to describe *schema:Product* records on websites newly annotating data using Microdata in 2013, all websites from 2013 and all websites from 2012 as well as the difference between the new websites and the all websites of 2012 and 2013. Outstanding differences are marked in bold.

	Property	New PLDs #	New PLDs %	PLDs'13 %	Change in %	PLDs'12 %	Change in %
1	s:Product/name	25 679	91.07	89.62	1.62	86.34	5.48
2	s:Product/description	19 977	70.85	67.45	5.03	61.99	14.29
3	s:Product/image	19 037	67.51	61.93	**9.02**	48.72	**38.58**
4	s:Product/offers	18 179	64.47	58.68	**9.86**	45.42	**41.94**
5	s:Offer/price	16 829	59.68	54.55	**9.41**	41.50	**43.81**
6	s:Offer/availability	11 977	42.47	37.40	**13.58**	10.29	**312.63**
7	s:AggregateRating	7 809	27.69	30.25	−8.45	25.93	6.79
8	s:Product/aggregateRating	7 664	27.18	29.26	−7.12	11.87	**128.96**
9	s:AggregateRating/ratingValue	7 469	26.49	28.95	−8.50	24.02	10.28
10	s:Offer/priceCurrency	6 934	24.59	24.28	1.29	9.63	**155.31**
11	s:Product/url	5 897	20.91	21.17	−1.20	12.90	62.11
12	s:Product/manufacturer	5 671	20.11	14.85	**35.44**	1.98	**915.47**
13	s:AggregateRating/reviewCount	5 662	20.08	20.94	−4.11	8.06	**149.11**
14	s:Product/productID	3 983	14.13	13.11	7.76	10.52	34.24
15	s:AggregateRating/bestRating	3 089	10.95	13.87	−21.01	16.10	−31.97
16	s:Product/brand	2 959	10.49	10.43	0.65	11.94	−12.09
17	s:Offer/itemCondition	2 659	9.43	6.86	**37.43**	2.16	**337.56**
18	s:AggregateRating/ratingCount	2 651	9.40	12.37	−24.01	16.21	−41.99
19	dv:Breadcrumb/url	2 131	7.56	7.73	−2.26	10.64	−28.99
20	dv:Breadcrumb/title	2 124	7.53	7.67	−1.82	10.63	−29.15

fashion compared to our datasets which can only contain data from HTML pages included in the Common Crawl.

The big search engine companies Google, Yahoo!, Microsoft and Yandex extract Microdata, RDFa, and Microformats data from their Web crawls but, for economic reasons, do not provide public access to the resulting datasets. Although they have published a number of studies about the deployment of the markup languages: Mika and Potter analyze the adoption of the languages based on Web crawls from the Bing search engine dating from 2011 and 2012 [10,11]. Guha presented an updated analysis of the deployment of Microdata with a special focus on the *Schema.org* vocabulary at the LDOW 2014 workshop [5].

10 Conclusion

This paper has presented a series of publicly accessible Microdata, RDFa, Microformats datasets that we have extracted from three large Web corpora dating from 2010, 2012 and 2013. The extracted datasets show that all three markup standards are used by hundreds of thousands of websites. Comparing the 2012 and 2013 datasets reveals that the number of websites using Microdata has grown by more than factor four in just one year. Altogether, the extracted datasets consist of almost 30 billion RDF quads and contain large quantities of product, review, address, blog post, people, organization, event, and cooking recipe data. As far as we know, the WebDataCommons datasets are the largest publicly accessible datasets of this kind.

We believe that the data will be useful for various applications such as building product catalogs, address databases or event and cooking websites. The data also

Table 10. Most frequently used Microformats classes within the 2013 dataset sorted by the number of websites using the class, including the total number of records in 2013 as well as the number of websites using the class in 2012

		2013			2012		
		Records	PLDs		Records	PLDs	
	Class	# (in k)	#	%	# (in k)	#	%
1	hCard:VCard	787 859	994 829	89.14	525 300 858	1 511 467	84.03
2	hCard:Organization	126 356	119 049	10.67	62 880 238	195 493	10.87
3	geo:Location	75 945	23 044	2.06	13 206 248	48 415	2.69
4	hCalendar:vcalendar	4 173	20 981	1.88	3 883 524	37 620	2.09
5	hCalendar:Vevent	37 989	17 633	1.58	28 737 655	36 349	2.02
6	hReview:Review	19 734	12 880	1.15	27 781 420	20 781	1.16
7	hRecipe:Recipe	1 009	3 530	0.32	1 260 116	3 281	0.22
8	hListing:Lister	9 016	2 584	0.23	9 992 047	4 030	0.22
9	hListing:Listing	9 016	2 584	0.23	9 992 047	4 030	0.18
10	hRecipe:Ingredient	6 825	2 524	0.23	8 405 151	2 658	0.16
11	hListing:Item	1 656	1 793	0.16	5 236 418	2 957	0.15
12	hRecipe:Duration	344	1 044	0.09	341 601	1 323	0.07
13	hRecipe:Nutrition	399	446	0.04	1 688 412	818	0.05
14	species:species	37	109	0.01	82 610	91	0.01
15	species:Genus	21	74	0.01	40 589	61	0.00
16	species:Family	20	72	0.01	40 651	60	0.00
17	species:Kingdom	19	72	0.01	40 833	59	0.00
18	species:Order	20	70	0.01	40 462	59	0.00

constitutes a valuable source of evaluation data for testing methods from various research areas. For evaluation purposes, the amount of data contained in the datasets should be large and representative enough. For commercial purposes, it has to be kept in mind that the Common Crawl only contains a subset of the pages from each website. Thus, the extracted datasets can also only contain a subset of the Microdata, RDFa, Microformats annotations offered by each website and should thus rather be used to identity seeds for more complete directed crawls. Before Microdata, RDFa, Microformats data can be used in application settings, several challenges need to be addressed:

Information Extraction: Most entities are only marked up with a relatively small number of properties and these properties tend to be rather generic, such as name or description properties, leading to rather flat records. It is thus often necessary to apply further information extraction methods to the property values in order to reach more fine grained data structures that allow the application of more sophisticated data integration and cleansing methods [12].

Identity Resolution: The data hardly contains entity identifiers, such as ISBN EAN numbers, which would make it easy to identity records from different websites that described the same entity. Instead, applications that want to deduplicate data from multiple websites need to match the entity descriptions published by the sites. An example of how such an identity resolution heuristic is applied to Microdata product records is given in [12].

Data Quality Assessment: As the Web is an open and unrestricted information environment, web data might be outdated or simply wrong. Thus, before data is used in an application context its quality should be assessed based on its content as well as its provenance. An interesting identity resolution

and data quality assessment challenge is for instance given by the Microdata address data: Which of the provided addresses is the current address of a company? How to determine this address given that many yellow pages websites copy from each other and simple voting thus does not work?

We believe that the adoption of the Microdata, RDFa, Microformats standards by hundreds of thousands of websites provides a huge potential for using Web data within various applications. On the other hand, it also raises tough challenges concerning the integration and cleansing of the data. By providing the WebDataCommons dataset series, we hope to contribute to addressing these challenges and to lift the potential of the data.

Acknowledgement. The extraction of the datasets from the Common Crawl was in part supported by the FP7-ICT projects PlanetData (GA 247641), and LOD2 (GA 257943) and by an Amazon Web Services in Education Grant award. We would like to thank the Common Crawl foundation for publishing the web crawls. We also thank the Any23 team for their great parsing framework as well as Hannes Mühleisen for his initial work on the Web Data Commons extraction framework.

References

1. Ben Adida and Mark Birbeck. RDFa primer - bridging the human and data webs - W3C recommendation (2008), http://www.w3.org/TR/xhtml-rdfa-primer/
2. Bizer, C., Eckert, K., Meusel, R., Mühleisen, H., Schuhmacher, M., Völker, J.: Deployment of rdfa, microdata, and microformats on the web a quantitative analysis. In: Alani, H., et al. (eds.) ISWC 2013, Part II. LNCS, vol. 8219, pp. 17–32. Springer, Heidelberg (2013)
3. Goel, K., Guha, R.V., Hansson, O.: Introducing rich snippets (2009), http://googlewebmastercentral.blogspot.de/2009/05/introducing-rich-snippets.html
4. Guha, R.V.: Schema.org support for job postings (2011), http://blog.schema.org/2011/11/schemaorg-support-for-job-postings.html
5. Guha, R.V.: Schema.org update (April 2014), http://events.linkeddata.org/ldow2014/slides/ldow2014_keynote_guha_schema_org.pdf
6. Haas, K., Mika, P., Tarjan, P., Blanco, R.: Enhanced results for web search. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, pp. 725–734. ACM, New York (2011)
7. Hickson, I.: HTML Microdata, Working Draft (2011), http://www.w3.org/TR/microdata/
8. Lindahl, G.: Blekko donates search data to common crawl (December 2012), http://blog.blekko.com/2012/12/17/common-crawl-donation/
9. Meusel, R., Vigna, S., Lehmberg, O., Bizer, C.: Graph structure in the web - revisited: a trick of the heavy tail. In: Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion, International World Wide Web Conferences Steering, pp. 427–432 (2014)

10. Mika, P.: Microformats and RDFa deployment across the Web (2011), http://tripletalk.wordpress.com/2011/01/25/rdfa-deployment-across-the-web/
11. Mika, P., Potter, T.: Metadata statistics for a large web corpus. In: LDOW 2012: Linked Data on the Web, CEUR Workshop Proceedings, vol. 937, CEUR-ws.org (2012)
12. Petrovski, P., Bryl, V., Bizer, C.: Integrating product data from websites offering microdata markup. In: 4th Workshop on Data Extraction and Object Search, DEOS 2014 (2014)