Metadata Statistics for a Large Web Corpus

Peter Mika
Yahoo! Research
Diagonal 177
Barcelona, Spain
pmika@yahoo-inc.com

Tim Potter
Yahoo! Research
Diagonal 177
Barcelona, Spain
tep@yahoo-inc.com

ABSTRACT

We provide an analysis of the adoption of metadata standards on the Web based a large crawl of the Web. In particular, we look at what forms of syntax and vocabularies publishers are using to mark up data inside HTML pages. We also describe the process that we have followed and the difficulties involved in web data extraction.

1. INTRODUCTION

Embedding metadata inside HTML pages is one of the ways to publish structured data on the Web, often preferred by publishers and consumers over other methods of exposing structured data, such as publishing data feeds, SPARQL endpoints or RDF/XML documents. Publishers prefer this method due to the ease of implementation and maintenance: since most webpages are dynamically generated, adding markup simply requires extending the template that produces the pages. Consumers such as search engines are already accustomed to processing HTML and extraction fits naturally in their processing pipelines. The close coupling of the raw data and the HTML presentation of the data has other advantages, among others it makes sure that the the raw data and the end-user presentation show the same.

In this paper, we describe the method by which we extracted metadata from a large web corpus and present some statistics. Results from similar experiments have been already published, so we also discuss the difficulty in comparing numbers across the various studies.

2. RELATED WORK

Previous studies have reported results on the usage of embedded metadata, including Bizer et al. at `http://www.webdatacommons.org/`. We also published an earlier analysis on a different corpus collected by Yahoo! Search [1]. There are a number of factors that complicate the comparison of results. First, different studies use different web corpora. Our earlier study used a corpus collected by Yahoo!'s web crawler, while the current study uses a dataset collected by the Bing crawler. Bizer et al. analyze the data collected by `http://www.commoncrawl.org`, which has the obvious advantage that it is publicly available. Second, the extraction methods may differ. For example, there are a multitude of microformats (one for each object type) and although most search engines and extraction libraries support the popular ones, different processors may recognize a different subset. Unlike the specifications of microdata and RDFa published by the RDFa, the microformat specifications are also rather informal and thus different processors may extract different information from the same page. Further, even if the same information is extracted, the conversion of this information to RDF may differ across implementations. Third, different extractors may be lenient in accepting particular mistakes in the markup, leading to more or less information extracted.

[1] `http://tripletalk.wordpress.com/2011/01/25/rdfa-deployment-across-the-web/`

LDOW2012, April 16, 2012, Lyon, France.

3. ANALYSIS

We take as our starting point a sufficiently large sample of the web crawl produced by Bing's web crawler during January, 2012. After retaining information resources with a content type that includes *text/html*, we get a data set of 3,230,928,609 records with only the three fields required for analysis, the URL of the page, the content type and the downloaded content. In case the crawler arrived to a page by following a (chain of) redirects, we considered the target of the redirect as the URL.

We perform our analysis in two steps. First, we use regular expression patterns to detect metadata in web pages. We use the same patterns proposed by Bizer et al., but we strengthen the pattern for detecting RDFa. In the form proposed by the authors it allows any page that contains *about* followed by whitespace and an equal sign;we limit this pattern to require that the equal sign be followed by whitespace and a single or double quote. We also introduce a new pattern to specifically detect webpages using the Open Graph Protocol Second, identified by the word *property* followed by optional whitespace, single or double quote, optional whitespace and *og:*. For this analysis, we filter out pages larger than 3MB and where the character set can not be identified. The total number of URLs in the output is thus slightly lower than in the input.

Table 1 shows the prevalence of each format both in terms of URLs that use that format, and in terms of effective top-level domains (eTLD), sometimes called pay-level domains

(PLD)[2]. For computing PLDs, we used the Guava library version 11.0.2. For a small number of URLs we failed to determine the PLD, e.g. because they contain an IP address instead of a domain name, but we believe this does not influence the results significantly.

In a second step, we actually extract RDFa data from these pages using the Any23 library (version 0.7) as suggested by Bizer et al., and using the same set of extractor plugins. We use this library with the default configuration except for setting metadata nesting[3] to *off*, because microformat extraction generates a substantial number of additional triples in the default setting. Before passing the content to Any23, we read the char set of the page from the content-type and recode the page content to UTF-8 (we exclude pages where the character set can not be identified). We also modify each input page that we expect to contain OGP markup to define the *og* prefix. Without this, much of OGP data would not be extracted by Any23's RDFa parser and there is also no specific extractor for OGP data. To speed up the process of extraction, we exclude some extreme cases: webpages larger than 3 MB, pages , pages containing more than 200 VCard objects, and also pages where the result of the extraction exceeds 64 MB. We write the data in a quintet format: subject, predicate, object, context and the name of the extractor that produced that quad.

To read the data, we use the same NxParser library that we use to write the data. Unfortunately, there are invalid lines in the output that we are not able to read back (various exceptions reported by NxParser). Further, some input lines cause the parsing to enter an infinite loop. As a temporary measure until we find the source of these bugs, we run the parser in a separate thread and terminate this thread after 500ms. We also limit the size of each input line to 5KB and do not even attempt to parse lines longer than that. Due to these problems, we loose some data: the output contains 671,454,122 URLs compared to 973,539,519 URLs that we would expect to contain some data based on regular expressions. In total, we extract 17,443,606,947 triples. Tables reftbl:topsites-rdfa and 3 and 4 show the top 10 sites as measured by the number of triples using RDFa, microdata, or hcard, respectively. The number of triples is an aggregate that reflects both the number of indexed pages in the crawl (a proxy for the importance of the domain) and the amount of data published per page. Again, we note that these lists are not exclusive. For example, youtube.com uses both microformats, microdata and RDFa within the same pages.

In terms of vocabulary usage, we show the most commonly used namespaces in RDFa data in Table 5. We also show the most frequently used classes in terms of the number of URLs and PLDs in Table 6 and Table 7, respectively. We omit the http protocol identifier, because all namespaces start with this protocol identifier, except for a facebook namespace that appears with both http and https. The first table confirms that the vast majority of RDFa data on the Web is due to Facebook's OGP markup. Unfortunately, OGP does not always conform with the letter and intent of RDFa. For example, type information in OGP is given using the *og:type* predicate, and not the RDF built-in *rdf:type* predicate. This explains the difference between Table 5 vs Table 6 and Ta-

[2]`http://en.wikipedia.org/wiki/Public_Suffix_List`
[3]any23.extraction.metadata.nesting

Site	Triple count
facebook.com	1,739,664,342
tabelog.com	662,028,717
venere.com	366,531,732
yahoo.com	223,125,828
tripadvisor.co.uk	195,314,434
tripadvisor.it	183,603,052
tripadvisor.com	179,970,956
tripadvisor.fr	134,442,146
tripadvisor.jp	125,976,435
tripadvisor.es	124,845,123
tripadvisor.de	96,635,499
answers.com	86,721,016
myspace.com	79,984,056
tripadvisor.in	69,763,161
daodao.com	66,014,882
tripadvisor.com.tw	63,430,680
tripadvisor.ru	41,199,304
imdb.com	40,537,631
youtube.com	39,942,197
bestbuy.com	35,910,433

Table 2: Top sites by number of triples, RDFa only

Site	Triple count
myspace.com	133,287,800
yelp.com	94,149,823
bbb.org	85,225,323
imdb.com	37,925,513
thefreelibrary.com	37,208,120
powells.com	31,056,409
youtube.com	26,299,315
homefinder.com	25,118,391
reverbnation.com	20,331,369
kino-teatr.ru	15,550,954
eventful.com	15,078,003
cylex.de	14,288,282
goodreads.com	12,484,280
bandcamp.com	11,372,475
bizrate.com	10,716,450
businesswire.com	9,488,095
wat.tv	9,280,173
avvo.com	9,113,367
barnesandnoble.com	8,444,559
patch.com	8,157,515

Table 3: Top sites by number of triples, microdata only

Format	Abs URL	Pct URL	Abs PLD	Pct PLD
RDFa	795,081,604	25.08 %	1,306,827	4.04%
OGP	711,747,491	22.45 %	1,140,880	3.53%
microdata	226,913,004	7.16 %	93,463	0.29%
microformat	272,470,501	8.60 %	1,755,733	5.43%
XFN	35,344,618	4.27 %	1,700,377	5.26%
no data	2,196,204,478	69.29 %	30,809,476	95.27%

Table 1: Results from pattern-based analysis $N_{URL} = 3,169,743,997$**,** $N_{PLD} = 32,339,522$

Site	Triple count
yahoo.com	572,687,378
twitter.com	534,336,425
linkedin.com	252,481,792
yellowpages.com	97,624,187
tvtrip.com	53,746,582
youtube.com	43,330,641
myspace.com	41,110,226
nii.ac.jp	40,752,988
nj.com	38,202,997
patch.com	38,003,049
chow.com	37,705,040
minecraftforum.net	35,891,626
oregonlive.com	33,159,011
everycarlisted.com	32,75,0040
nydailynews.com	32,211,122
last.fm	30,302,919
citysearch.com	28,444,466
washingtonpost.com	27,926,328
nieuwsblad.be	27,497,607
cleveland.com	26,998,847

Table 4: Top sites by number of triples, hcard only

ble 7: most OGP data does not define instances of any RDF class. As already mentioned above, most users of OGP also ignore the declaration of the *og* prefix (a problem we deal with in the extraction) and we can also see a number of variations to the current standard namespace (a problem we have not dealt with). Further, OGP assigns additional meaning to the RDFa syntax that is not reflected in the RDFa standard. As an example, the order in which triples are written on the page matters in OGP, but not in RDFa. For all these reasons, we believe that Any23 should be extended with a specific processor for OGP markup that is able to deal with these peculiarities.

Besides OGP, a smaller amount of data can be attributed to efforts by Google's Rich Snippet program and Yahoo's retired SearchMonkey program. Social markup in the form of FOAF and SIOC is also present in a large number of domains as shown in Table 7. The fact that these vocabularies do not show up as prominently in Table 6 means that they are used more in the less deeply crawled part of the web.

For microdata, we only list the top namespaces in Table 8 and Table 9, because Any23's microdata extractor incorporates the class name into the namespace. In microdata, only two vocabularies (*schema.org* and Google's *data-vocabulary.org*) have gained significant traction so far, and the latter is expected to be replaced by the former.

It holds for both RDFa and microdata that the types of objects that are marked up is biased by the use case of search engine optimization, i.e. site owners prefer to mark up data that is used by the search engines to enrich search result presentation (e.g reviews, business listings). Schemas for these types of objects have also existed longer. We also observe a natural preference to mark up simple types of objects (e.g. breadcrumbs), though we did not formally investigate the relationship between the complexity of markup and its adoption.

4. CONCLUSIONS

We presented metadata statistics from the analysis of a large, recent sample of the Web, which has been extracted from the crawl of a search engine and therefore provides a search-engine centric view on the Web. Current web search engines are biased toward authoritative, head sites with valuable textual content, and are not specifically looking for data on the Web. We expect that a search engine specifically built for data would give less weight to authority and textual content and perform deeper crawling on sites that provide large and valuable data, by some measure of quantity and quality.

Nonetheless, our work shows an impressive progress in the adoption of markup on the Web with over 30% of our collection containing some microformat, RDFa or microdata markup. Microformats and RDFa are the most popular choices of syntax. The level of microformats usage seems to be flat, while RDFa adoption has grown significantly compared to previous studies. This is due almost exclusively to OGP markup, though there is a variety of usage in the long tail, in particular social vocabularies. On the other hand, the adoption of microdata is driven so far only by the success of *schema.org*.

There is significant future work to be done in order to evaluate the quality and practical usefulness of data embedded in HTML, with respect to some existing or novel tasks. In previous work, we have looked at the extent to which embedded metadata could be used to enrich web search results [1], but data on the Web is likely to be useful in a much broader array of applications.

5. REFERENCES

[1] K. Haas, P. Mika, P. Tarjan, and R. Blanco. Enhanced results for web search. In W.-Y. Ma, J.-Y. Nie, R. A. Baeza-Yates, T.-S. Chua, and W. B. Croft, editors, *SIGIR*, pages 725–734. ACM, 2011.

Namespace	URLs
ogp.me/ns#	493,443,016
www.facebook.com/2008/	150,246,016
www.w3.org/1999/02/22-rdf-syntax-ns#	26,402,165
rdf.data-vocabulary.org/#	19,413,470
purl.org/dc/terms/	16,424,800
https://www.facebook.com/2008/	7,472,815
mixi-platform.com/ns#	6,323,861
ogp.me/ns/fb#	4,636,260
creativecommons.org/ns#	4,622,272
www.w3.org/2006/vcard/ns#	4,205,037
http://	3,881,321
http://www.facebook.com/	3,126,045
http://www.w3.org/2000/01/rdf-schema#	3,042,839
http://developers.facebook.com/schema/	2,720,567
http://search.yahoo.com/searchmonkey/commerce/	2,664,743
http://purl.org/dc/elements/1.1/	2,642,796
http://opengraphprotocol.org/schema/	2,293,024
http://search.yahoo.com/searchmonkey/media/	2,095,577
http://oexchange.org/spec/0.8/rel/	2,034,467
http://xmlns.com/foaf/0.1/	1,837,749

Table 5: Top namespaces in RDFa as measured by the number of URLs

Class	URLs
rdf.data-vocabulary.org/#Breadcrumb	11,336,922
rdf.data-vocabulary.org/#Review-aggregate	5,571,178
rdf.data-vocabulary.org/#Organization	3,678,229
www.w3.org/2006/vcard/ns#VCard	2,858,916
search.yahoo.com/searchmonkey/commerce/Business	2,727,213
rdf.data-vocabulary.org/#Review	1,980,811
rdf.data-vocabulary.org/#Rating	1,714,996
rdf.data-vocabulary.org/#review-aggregate	1,453,439
xmlns.com/foaf/0.1/Image	1,446,290
search.yahoo.com/searchmonkey/product/Product	1,202,002
http://rdf.data-vocabulary.org/#Address	1,087,380
http://www.purl.org/stuff/rev#Review	746,858
http://rdf.data-vocabulary.org/#Product	673,079
http://purl.org/goodrelations/v1#UnitPriceSpecification	648,598
http://purl.org/goodrelations/v1#Offering	599,703
http://xmlns.com/foaf/0.1/Agent	517,089
http://xmlns.com/foaf/0.1/Document	441,694
http://www.w3.org/2004/02/skos/core#Concept	406,776
http://xmlns.com/foaf/0.1/Group	369,176
http://rdfs.org/sioc/ns#Item	363,308

Table 6: Top classes in RDFa as measured by the number of URLs with at least one instance

Class	PLDs
xmlns.com/foaf/0.1/Image	30,903
xmlns.com/foaf/0.1/Document	25,090
rdfs.org/sioc/ns#Item	19,583
rdfs.org/sioc/ns#UserAccount	15,058
www.w3.org/2004/02/skos/core#Concept	9,757
rdf.data-vocabulary.org/#Breadcrumb	5,427
rdfs.org/sioc/ns#Post	5,342
rdf.data-vocabulary.org/#Review-aggregate	3,307
rdfs.org/sioc/types#BlogPost	2,970
rdfs.org/sioc/types#Comment	2,695
http://rdf.data-vocabulary.org/#Rating	2,114
http://rdf.data-vocabulary.org/#Organization	1,759
http://www.w3.org/2006/vcard/ns#Address	1,655
http://purl.org/goodrelations/v1#BusinessEntity	1,608
http://purl.org/goodrelations/v1#UnitPriceSpecification	1,385
http://rdf.data-vocabulary.org/#Review	1,294
http://rdf.data-vocabulary.org/#Product	1,246
http://purl.org/goodrelations/v1#QuantitativeValue	1,051
http://rdf.data-vocabulary.org/#Address	932
http://purl.org/goodrelations/v1#Offering	787

Table 7: Top classes in RDFa as measured by the number of PLDs with at least one instance

Namespace	URLs
www.w3.org/1999/xhtml/microdata#	67,087,467
www.w3.org/1999/02/22-rdf-syntax-ns#	66,745,726
purl.org/dc/terms/	46,675,266
data-vocabulary.org/Breadcrumb/	19,368,347
schema.org/MusicGroup/	6,699,903
schema.org/MusicRecording/	6,591,236
schema.org/Person/	4,650,659
schema.org/Product/	3,667,023
schema.org/VideoObject/	3,228,156
http://schema.org/Article/	3,052,457
http://schema.org/WebPage/	2,928,410
http://data-vocabulary.org/Product/	2,742,977
http://schema.org/PostalAddress/	2,736,213
http://schema.org/Offer/	2,553,617
http://data-vocabulary.org/Review-aggregate/	2,152,533
http://schema.org/AggregateRating/	2,048,232
http://schema.org/LocalBusiness/	2,043,005
http://schema.org/Organization/	1,640,501
http://data-vocabulary.org/Offer/	1,628,027
http://schema.org/Review/	1,281,548

Table 8: Top namespaces in microdata as measured by the number of URLs

Namespace	PLDs
data-vocabulary.org/Breadcrumb	14,623
schema.org/PostalAddress	11,476
schema.org/LocalBusiness	8,820
schema.org/Product	6,817
data-vocabulary.org/Organization	3,765
schema.org/Offer	3,654
schema.org/Organization	3,614
data-vocabulary.org/Address	3,529
schema.org/Article	3,283
schema.org/MusicGroup	3,253
http://schema.org/MusicAlbum	2,974
http://www.schema.org/MusicRecording	2,941
http://schema.org/Person	2,676
http://data-vocabulary.org/Product	2,596
http://data-vocabulary.org/Review-aggregate	2,450
http://schema.org/AggregateRating	2,380
http://schema.org/WebPage	2,132
http://data-vocabulary.org/Rating	1,947
http://schema.org/GeoCoordinates	1,651
http://schema.org/Place	1,634

Table 9: Top namespaces in microdata as measured by the number of PLDs