# Towards Crawling the Web for Structured Data: Pitfalls of Common Crawl for E-Commerce

Alex Stolz and Martin Hepp

Universitaet der Bundeswehr Munich, D-85579 Neubiberg, Germany
{alex.stolz,martin.hepp}@unibw.de

**Abstract.** In the recent years, the publication of structured data inside HTML content of Web sites has become a mainstream feature of commercial Web sites. In particular, e-commerce sites have started to add RDFa or Microdata markup based on schema.org and GoodRelations vocabularies. For many potential usages of this huge body of data, we need to crawl the sites and extract the data from the markup. Unfortunately, a lot of markup resides in very deep branches of the sites, namely in the product detail pages. Such pages are difficult to crawl because of their sheer number and because they often lack links pointing to them. In this paper, we present a small-sized experiment where we compare the Web pages from a popular Web crawler, Common Crawl, with the URLs in sitemap files of respective Web sites. We show that Common Crawl lacks a major share of the product detail pages that hold a significant part of the data, and that an approach as simple as a sitemap crawl yields much more product pages. Based on our findings, we conclude that a rethinking of state-of-the-art crawling strategies is necessary in order to cater for e-commerce scenarios.

## 1 Introduction

Nowadays, evermore vendors are offering and selling their products on the Web. The spectrum of online vendors ranges from a few very large retailers like Amazon and BestBuy featuring an ample amount of goods to many small Web shops with reasonable assortments. Consequently, the amount of product information available online is constantly growing.

While product descriptions were mostly unstructured in the past, the situation has meanwhile changed. In the last few years, numerous Web shops have increasingly started to expose product offers using structured data markup embedded as RDFa, Microdata, or microformats in HTML pages (cf. [1,2]). Standard e-commerce Web vocabularies like GoodRelations or schema.org typically complement these data formats (RDFa and Microdata[1]) by semantically describing products and their commercial properties. To some extent, the semantic annotations of product offers on the Web have been promoted by search engine

[1] Unlike for pure syntactical formats like RDFa and Microdata, the class and property names in microformats already imply semantics.

operators that offer tangible benefits and incentives for Web shop owners. In particular, they claim that additional data granularity can increase the visibility of single Web pages, for example in the form of *rich snippets* (or *rich captions*, in Bing terminology) displayed on search engine result pages [3,4]. Furthermore, structured data – and structured product data in particular – can provide useful relevance signals that search engines can take into account for their ranking algorithms as for the delivery of more appropriate search results. In essence, structured product data opens up novel opportunities for sophisticated use cases such as deep product comparison at Web scale.

Unfortunately, for regular data consumers other than search engines like Google, Bing, Yahoo!, or Yandex[2], it is difficult to make use of the wealth of product information on the Web. Web crawling is an open problem for it is known to be expensive due to the sheer size of the Web, and because it requires deep technical expertise. Hence, motivated by these challenges and the idea to provide easy and open access to periodic snapshots of industrial-strength Web crawls, the Common Crawl project was launched in 2007.

The Common Crawl corpus with its hundreds of terabytes' worth of gathered data constitutes an incredibly valuable resource for research experiments and application development. At almost no cost[3], it can offer unprecedented insights into several aspects of the Web. Among others, researchers found it to be useful to analyze the graph structure of the Web [5], or to support machine translation tasks (e.g. [6]). Moreover, it is possible to extract and analyze structured data on a large scale, as tackled by the Web Data Commons initiative [1].

One popular and persistent misconception about Common Crawl, however, is to think that it is truly representative for the Web as a whole. The FAQ section on `commoncrawl.org` is further contributing to this fallacy, as it literally terms Common Crawl as "a copy of the web" [7]. This statement is problematic though, because it can foster wrong decisions. In particular, Web-scale crawling is costly, that is why Web crawlers generally give precedence to a subset of pages, selected by popularity or relevance thresholds of Web shops. To prioritize the visiting of Web pages, they take into account sitemaps, page load time, or the link structure of Web pages (i.e. *PageRank* [8]). Common Crawl e.g., employs a customized PageRank algorithm and includes URLs donated by the former search engine Blekko[4]. Yet, a lot of the markup can be found in very deep branches of the Web sites, like product detail pages. Aforementioned crawling strategies miss out many of these long-tail product pages, which renders them inappropriate for commercial use cases. For a related discussion on the limited coverage of Common Crawl with respect to structured e-commerce data, we refer to a W3C mailing list thread from 2012[5].

[2] Shop owners often proactively direct popular search engines to their Web pages.
[3] The only costs that incur are those for storage and processing of the data.
[4] http://www.slideshare.net/davelester/introduction-to-common-crawl
[5] http://lists.w3.org/Archives/Public/public-vocabs/2012Mar/0095.html and http://lists.w3.org/Archives/Public/public-vocabs/2012Apr/0016.html

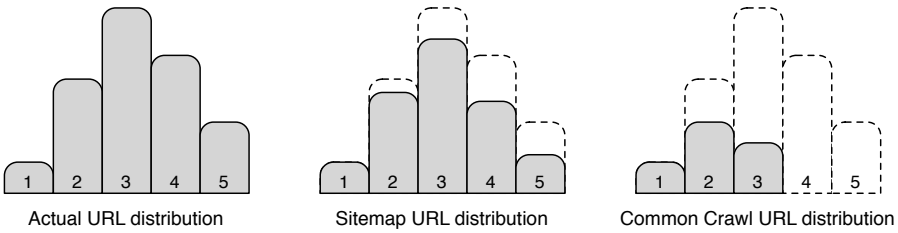



**Fig. 1.** Our initial assumption about the URL coverage at varying Web site levels

The main motivation for this paper was to answer the question whether Common Crawl is complete enough for use cases around structured e-commerce data. For e-commerce applications e.g., it is crucial to be able to rely on a solid and comprehensive product information database, which we think Common Crawl cannot offer to date. Even if it might not be very surprising that Common Crawl misses out a considerable amount of product data in commercial Web sites, it remains so far unclear by how much.

In this paper, we investigate the representativeness of Common Crawl with respect to e-commerce on the Web. For this purpose, we rely on a random sample of product-related Web sites from Common Crawl, for which we measure the coverage of (1) Web pages in general, and (2) Web pages with structured product offer data. We further analyze the crawl depth of Common Crawl. As our baseline, we use sitemaps to approximate the actual URL distribution of Web sites, as depicted in Fig. 1.

The rest of this paper is structured as follows: Section 2 reviews important related work; in Sect. 3, we outline our method, the relevant data sources, and design decisions that we made; Section 4 reports the results of the analysis of Common Crawl; in Sect. 5, we discuss our findings and emergent shortcomings; and finally, Sect. 6 concludes our paper.

## 2 Related Work

Considerable literature work deals with crawling structured data from the Web. In contrast to conventional Web crawling of textual content, structured data poses special challenges on the organization and performance of crawling and indexing tasks [9]. The work in [9] suggests a pipelined crawling and indexing architecture for the Semantic Web. In [10], the authors propose a learning-based, focused crawling strategy for structured data on the Web. The approach uses an online classifier and performs a bandit-based selection process of URLs, i.e. it determines a trade-off between exploration and exploitation based on the page context and by incorporating feedback from previously discovered metadata. The Web Data Commons [1] initiative reuses a Web crawl provided by Common Crawl to extract structured data.

Surprisingly, none of the aforementioned approaches address the question where the structured data actually resides in a Web site, albeit this information can be useful for tuning the crawling process. [11] provide a study on the distribution of structured content for information extraction on the Web. They show that in order to capture a reasonable amount of the data of a domain, it is crucial to regard the long tail of Web sites. On a more general scope, the authors of [12] estimate that a crawler is able to reach a significant portion of the Web pages visited by users with only following three to five steps from the start page.

## 3 Methodology and Data

In this paper, we address the problem of crawling e-commerce data from the Web, and in particular we focus on caveats of using Common Crawl. This section describes the data sources and the methodology for the inspection of the Common Crawl relative to e-commerce.

### 3.1 Datasets and APIs

Each of the Common Crawl corpora represents a huge body of crawled Web pages which to process and analyze in detail is not a straightforward task. The December 2014 Common Crawl corpus has a compressed file size of 160 terabytes and is reported to contain 15.7 million Web sites and a total of over two billion Web pages [13]. To cope with this enormous amount of data, we drew on derived datasets and APIs.

**Web Data Commons.** The Web Data Commons dataset series contains all structured data extracted from the various Common Crawl corpora [1]. Currently, the extraction process considers the data formats Microdata, RDFa, and microformats. In addition to the publishing of the comprehensive datasets, the Web Data Commons initiative also provides class-specific subsets to the most prominent schema.org classes. We use the subset for *s:Offer* from the December 2014 Common Crawl corpus, which is comprised in a 48 gigabyte gzip-compressed N-Quads file[6].

**Common Crawl Index Query API.** In order to access the huge Common Crawl corpora, it is no longer necessary to get hold of the entire datasets. In 2015, the Common Crawl Index was announced, which offers an index query API to look up entries by URL (possibly with wildcards). The Common Crawl Index for the December 2014 crawl snapshot is available at `http://index.commoncrawl.org/CC-MAIN-2014-52`. For the Python programming language, there further exists an API client to conveniently access the Common Crawl indices[7].

[6] `http://webdatacommons.org/structureddata/2014-12/stats/schema_org_subsets.html`

[7] `https://github.com/ikreymer/cdx-index-client`

**Sitemaps.** In addition to the data sources related to Common Crawl, we obtained Web page URLs and corresponding metadata from the XML sitemaps of selected Web sites.

### 3.2 Method

In any e-commerce Web site, the most distinctive entity type is the product offer, which describes the commercial details and terms and conditions related to the transfer of property rights for a product item. Product offers are represented in schema.org by the class *s:Offer*. Within the scope of this research, we used the respective class-specific subset of Web Data Commons that was extracted from the December 2014 snapshot of Common Crawl.

We extracted all available domains from the Web Data Commons subset related to *s:Offer*. From the 109651 domains obtained this way, we drew a random sample of 100 Web sites. For each of these Web sites, we (a) captured the available URLs in the Common Crawl Index, and (b) fetched their XML sitemaps (if existing) to collect the contained URLs. In order to locate the sitemaps, we tested whether they are referenced from a *robots.txt* file or, alternatively, placed in the root directory of the Web server – which is regarded as the simplest and default method. Then we (recursively) iterated over these sitemaps (or sitemap indices) and extracted the Web page URLs.

After that, we matched the Web pages found in the Web Data Commons subset to the URLs found in the respective sitemaps that contain structured product offer data. For this purpose, we conducted a deep crawl based on the URL collection in the sitemaps. We limited the number of URLs to 10000 per sitemap, otherwise the crawling process would have been overly resource-intensive.

### 3.3 Design Decisions

To find out how deeply the Common Crawl spider is reaching within Web site structures, we rely on a basic property of URLs, which is that URL paths are organized hierarchically [14, Section 1.2.3]. For all URLs in both datasets, we thus analyzed their URL structures. In particular, we classified them by matching their URL paths against the following string patterns

```
lev0: /{0}
lev1: /{0}/{1}
lev2: /{0}/{1}/{2}
...
```

The placeholders (indicated by numeric values between curly braces) describe character sequences of arbitrary length. The sequence may contain any character but the slash symbol. In order to pertain to a specific category, the patterns must be matched exactly. The URL scheme together with the authority (often the host) constitute level 0 URLs (labeled as category "lev0") because the URL path is empty, e.g. `http://www.example.org/`. Other URLs with a path in the root directory also belong to level 0, e.g. `http://www.example.org/index.html`

or http://www.example.org/index, whereas http://www.example.org/index/ would be classified as a level 1 ("lev1") URL.

Prior to the analysis, the data needed to go through some pre-processing steps. In particular, we dropped duplicate URLs and then, for the purpose of generating statistics about the URL depth, we truncated query strings from URLs. The execution order of these two steps was essential, because otherwise we would have inadvertently discarded Web pages that only differ based on query string parameters. The reason we abstract from URLs with query string parameters is because we are only interested in the hierarchical characteristic of URL identifiers. This way we prevent the false classification of edge cases like http://www.example.org/?path=/some/path/, which should be treated equivalent to http://www.example.org/.

The random sample of product domains that we obtained was initially not very useful. We noticed some unexpected difficulties with various domains, e.g. many Web sites are part of large hosting platforms that are divided into several subdomains (e.g. masstube.uptodown.com and www.uptodown.com). As we could observe, the sitemap files of these domains often point at the main Web site (i.e. the subdomain starting with "www"). We thus decided to only consider Web sites starting with "www", by which we could reduce the problem considerably. In addition, it has the nice side-effect that possible www-subdomains are consolidated with their non-www-counterparts.

## 4 Analysis of Common Crawl Dataset for E-Commerce

From the 100 domains that we have examined, 68 featured a sitemap (15 a sitemap index)[8] according to the search criteria mentioned in the previous section. Out of these 68 remaining Web sites, we crawled all but *seven* domains that had no more than 10000 URLs. We then compared the numbers related to pages (URLs) with and without structured product offer data in the 61 domains with those from the Common Crawl corpus from December 2014. The datasets used for the analysis are published online[9].

### 4.1 Dataset Statistics

Table 1 gives an overview of the URLs and structured product offer data contained in each investigated Web site[10]. Columns 2-3 denote the number of URLs and the number of Web pages for which we could detect structured product offer data in the Web Data Commons dataset. By comparison, columns 4-5 report the exact same information about sitemaps, namely the number of URLs and the number of Web pages that contain structured product offer data according to a sitemap crawl.

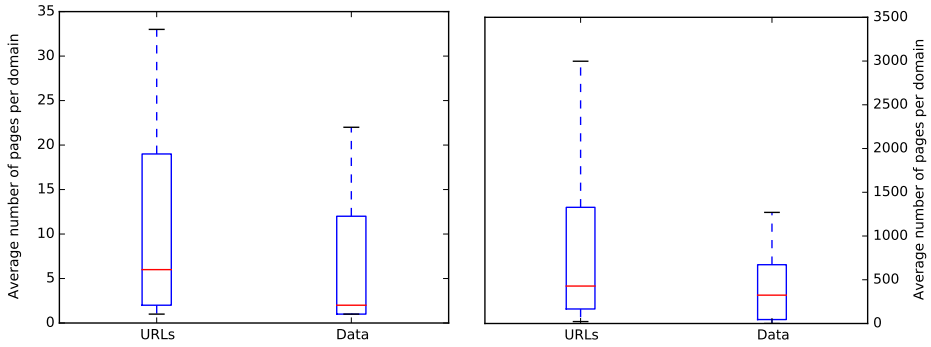
[8] Download date as of June 23, 2015

[9] http://www.ebusiness-unibw.org/resources/cold2015/data/

[10] We obtained the statistics using a string match "schema.org/Offer" against the HTML content, which is a simple heuristic to match schema.org in Microdata.

**Table 1.** Comparison of Web sites in Common Crawl with sitemap statistics

| | Common Crawl | | Sitemap | |
|---|---|---|---|---|
| | URLs | Data | URLs | Data |
| 01 www.abloomaboveflorist.com | 1 | 1 | 424 | 409 |
| 02 www.acuderm.com | 24 | 22 | 794 | 661 |
| 03 www.antik-zentrum-alling.de | 1 | 1 | 1218 | 1163 |
| 04 www.askariel.com | 53 | 14 | 107 | 43 |
| 05 www.bellyarmor.com | 12 | 1 | 40 | 8 |
| ... | | | | |
| 21 **www.hamiltonparkhotel.com** | 9996 | 7 | 63 | 21 |
| ... | | | | |
| 57 www.therealthingonline.co.za | 2 | 2 | 47 | 24 |
| 58 www.waterbedbargains.com | 27 | 21 | 315 | 299 |
| 59 www.wedoboxes.co.uk | 2 | 1 | 66 | 50 |
| 60 www.windberflorist.com | 1 | 1 | 431 | 416 |
| 61 www.xpradlo.sk | 1 | 1 | 1433 | 0 |




(a) Common Crawl (b) Sitemaps

**Fig. 2.** Average number of URLs and Web pages with *s:Offer* data (please note the different scales of the ordinate axes)

Figure 2 illustrates the boxplots corresponding to the data in Table 1. We cut off outliers from the boxplot in Fig. 2. There was e.g. one outlier in the Common Crawl which deviated significantly from the URLs available in the sitemap (see Table 1). While the site `www.hamiltonparkhotel.com` is represented by 9996 URLs in Common Crawl, its sitemap only contained 63 URLs. A closer look into the data revealed that Common Crawl collected plenty of deep Web links (9950 URLs) from an event calendar published on that site.

**Table 2.** Median and mean values of Web pages in Common Crawl and sitemaps, grouped by URL depth (for a subset of $n = 68$ sites)

| | **Common Crawl** | | **Sitemaps** | |
|---|---|---|---|---|
| | median | mean | median | mean |
| lev0 | 2 | 15.470588 | 15.5 | 800.588235 |
| lev1 | 1 | 205.602941 | 37.5 | 884.220588 |
| lev2 | 0 | 59.338235 | 17.0 | 1022.852941 |
| lev3 | 0 | 46.352941 | 0.0 | 209.073529 |
| lev4 | 0 | 22.323529 | 0.0 | 43.764706 |
| lev5 | 0 | 4.352941 | 0.0 | 13.852941 |
| lev6 | 0 | 0.735294 | 0.0 | 17.102941 |
| lev7 | 0 | 0.000000 | 0.0 | 0.014706 |
| lev8 | 0 | 0.382353 | 0.0 | 62.750000 |
| lev9 | 0 | 0.000000 | 0.0 | 0.014706 |
| lev10 | 0 | 0.647059 | 0.0 | 175.750000 |
| lev11 | 0 | 0.000000 | 0.0 | 0.014706 |

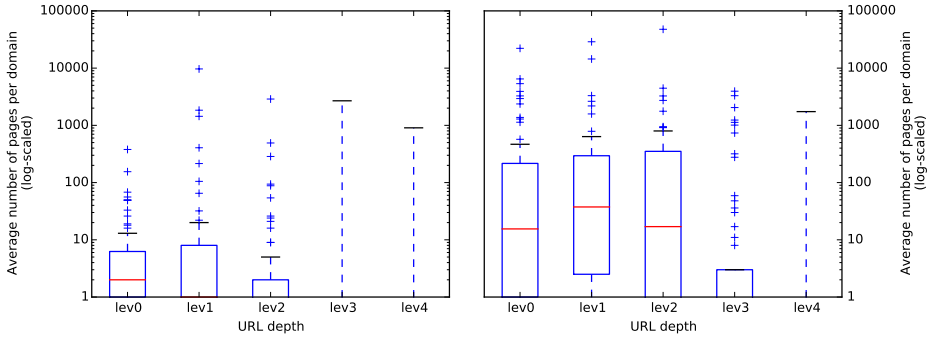
### 4.2 Coverage by URL Depth of Web Pages

In accordance with the assumption from Fig. 1, we compared the URLs in the Common Crawl corpus to the collection of URLs in the 68 sitemaps at varying URL depth levels. To accomplish that, we relied on the hierarchical part of the Web page URLs, as described in Sect. 3.2. Table 2 contrasts the domain-specific median and mean values of Common Crawl and sitemaps for the first twelve URL hierarchy levels. The median in the Common Crawl decreases across different levels, whereas for sitemaps there is first an increase from level 0 to level 1. Although the median value is then suddenly decreasing, it is still higher at level 2 than at level 0.

In Fig. 3, we show a boxplot of the data that was reported in Table 2. The figure graphically represents the distribution of URLs across Web sites for the first five URL depth levels.

### 4.3 Coverage by URL Depth of Web Pages with Structured Product Offer Data

In addition to measuring the coverage of URLs, we compared the Web pages that contain structured product offer data. The comparison was conducted among the 61 crawled sitemaps and the respective subset of Web Data Commons. Once again we did this comparison for the individual URL hierarchy levels. Table 3 outlines the median and mean values of URLs in sitemaps that contain structured product offer data, namely entities of type *s:Offer*. In comparison to Table 2, we added a rightmost column that shows by how much the mean values differ across Common Crawl and the sitemaps. E.g., the mean value between Common Crawl and the sitemaps differs by a factor of close to 67 at level 0 (the root directory). For level 1 it is 27.55, for level 2 53.22, and so on. This extrapolation factor gives




(a) Common Crawl (b) Sitemaps

**Fig. 3.** Web page frequency at the first five URL depth levels (for a subset of $n = 68$ sites)

**Table 3.** Median and mean values of Web pages with structured data markup for offers in Common Crawl and sitemaps, grouped by URL depth (for a subset of $n = 61$ sites)

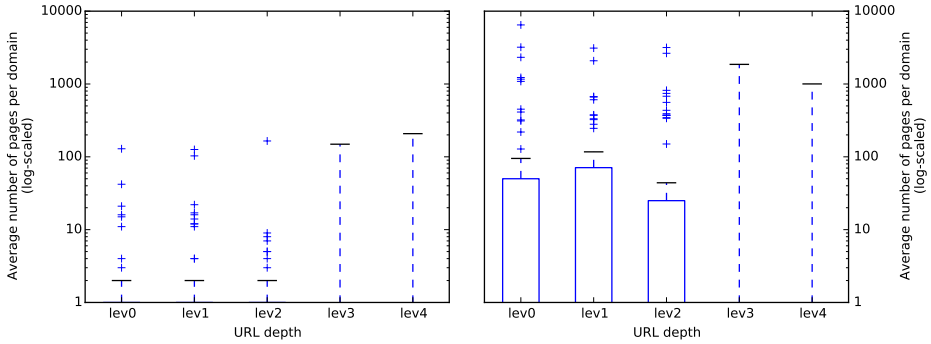
| | **Common Crawl** | | **Sitemaps** | | **Extrapolation** |
|---|---|---|---|---|---|
| | median | mean | median | mean | factor |
| lev0 | 0 | 4.360656 | 0 | 291.950820 | 66.95 |
| lev1 | 0 | 5.770492 | 0 | 158.983607 | 27.55 |
| lev2 | 0 | 3.557377 | 0 | 189.311475 | 53.22 |
| lev3 | 0 | 2.524590 | 0 | 71.918033 | 28.49 |
| lev4 | 0 | 3.409836 | 0 | 22.770492 | 6.68 |
| lev5 | 0 | 1.606557 | 0 | 3.737705 | 2.33 |
| lev6 | 0 | 0.245902 | 0 | 0.000000 | 0.00 |
| lev7 | 0 | 0.000000 | 0 | 0.016393 | inf |

an estimate of the missing coverage of structured e-commerce data in Common Crawl. In other words, if we would solely rely on sitemaps, then we could collect $x$ times as many Web pages with structured product offer data at a URL depth of $y$ than found in the Common Crawl dataset.

In Fig. 4, we show two boxplots for the URL collections in Common Crawl and in the sitemaps. They contrast the numbers of Web pages with structured product offer data at five different URL depth levels.

### 4.4 Evaluation of Coverage at URL Depth Levels

For the evaluation of our results, we analyzed by how much the coverage of URLs with structured product offer data deviates between the Common Crawl corpus and the URL collections from the sitemaps. More precisely, if we did encounter level 0 URLs in the Web Data Commons dataset, then we should also be able to find level 0 URLs within the sitemaps of the corresponding domains. Otherwise,




(a) Common Crawl (b) Sitemaps

**Fig. 4.** Web page frequency with *s:Offer* data at the first five URL depth levels (for a subset of $n = 61$ sites)

**Table 4.** Domain matches for Web pages with structured data markup for offers between Common Crawl and sitemaps

| | **Common Crawl** | **Sitemaps** | **Matches** |
|---|---|---|---|
| lev0 | 28 | 27 | 23 |
| lev1 | 20 | 30 | 18 |
| lev2 | 18 | 25 | 16 |
| lev3 | 6 | 10 | 4 |
| lev4 | 1 | 3 | 1 |

the averages reported in the results would be meaningless due to originating from distinct domains. We detected a considerable overlap between Common Crawl and sitemaps across various URL depth levels. The results are outlined in Table 4.

## 5 Discussion

In this section, we discuss the results of our analysis. Furthermore, we want to point at potential shortcomings of our approach.

Sitemaps of Web sites attain a better coverage of URLs than Common Crawl. A side-by-side comparison of the boxplots in Fig. 2 reveals differences between Common Crawl and sitemaps concerning (a) the average number of URLs in Web sites and (b) the average number of Web pages with product offers in Web sites. The two boxplots have a similar distribution, albeit the numbers differ significantly as the scale of the y-axis used for the sitemaps is 100 times larger than for Common Crawl. A comparison of the median values also indicates a larger share of Web pages with structured data markup for product offers in the sitemaps.

Common Crawl does not crawl as deep as the URL depth available in the sitemaps. Yet, Table 2 and Fig. 3 indicate that Common Crawl is well able to reach the deep pages within a Web site. However, in most cases it stops after the second URL depth level.

Against our expectations, a large amount of product offers is encountered at the upper three levels of the URL path rather than at the deeper hierarchy levels (see Fig. 4). The maximum URL depth in our datasets was twelve levels (see Table 2), although the deepest URL path with structured product offer data was eight (see Table 3). This might be due do different uses of hierarchical identifiers.

Interestingly, while the average number of structured product offers tends to decrease in the sitemaps, it stays relatively stable for the first seven levels in the Common Crawl, as shown in Table 3. This finding suggests that Common Crawl is able to enter Web sites at any URL depth levels, but instead of crawling additional URLs, it leaves Web sites prematurely – likely because of the prioritization strategy inherent to many Web crawlers.

In the course of this work, we have identified some possible limitations:

- There does not only exist *s:Offer* for entity types, but also concepts from other vocabularies and formats for annotating product offers on the Web (*gr:Offering* or microformats). We decided to rely on schema.org because it is by far the most prevalent Web vocabulary.
- Between the Common Crawl snapshot (December 2014) and the sitemap crawl (June 2015) was a time gap of half a year, which was unavoidable due to the lagged publication of the Web Data Commons extraction results.
- Sitemaps are useful for crawling large Web sites very quickly, but they are not necessarily complete as they are often webmaster-crafted. However, we do not see a disadvantage in here, rather would it in the best case further support our arguments.
- We compared the coverage of structured product offer data based on a sample of domains featuring sitemaps. Thus, our dataset could exhibit a potential bias towards higher-quality Web sites, i.e. those that are bigger in size or have more product detail pages. This issue needs further investigation.
- We assumed a hierarchical organization of URLs, despite some Web sites often use flat hierarchies for URL identifiers – or hierarchical identifiers in the absence of hierarchies, e.g. for tagging. Furthermore, the URL pattern used to classify URLs into various levels of URL depth is brittle because, depending on the server configuration, trailing slashes might be handled differently, e.g. `http://www.example.org` and `http://www.example.org/`.
- Finally, given the population size of 109651 Web sites, our sample size of 100 domains was relatively small which could be enhanced in future work.

## 6 Conclusion

In this paper, we have shown that the Common Crawl corpus lacks a large amount of product detail pages that hold a majority of the product data markup, which limits the usefulness of such a crawler for many e-commerce scenarios. We

provide evidence that the use of Common Crawl as the basis for large-scale data extraction, as done by the Web Data Commons initiative, yields a very incomplete body of data in that domain.

For a random sample of 100 Web sites from Common Crawl with structured product offer data, we have detected that the majority of them (i.e., 68) offer sitemap files. We have shown that Common Crawl covers only a small fraction of the URLs available in these sitemaps. Moreover, the amount of structured product offer data was generally lower for the Web sites considered in this work. Yet, the URL depth for sitemaps was only marginally higher than the crawl depth of Common Crawl.

Through the insights gained in this paper, we conclude that we need other crawling strategies that reach beyond what Common Crawl is currently able to offer for such purposes that need more than a more or less representative subset of all pages from considered Web sites. For example, to increase the overall coverage, we could combine Common Crawl with a deep sitemap crawl. Still, there remains the problem that a crawler like Common Crawl might miss the relevant Web pages that hold the product information within Web sites. Our findings presented in this paper apply to e-commerce sites in general as well as to e-commerce sites with structured markup. While we have not yet analyzed this for other domains or schema.org types, it is very likely that the same problems exist for other database-driven Web sites with a large number of entities in the backend database and little or no external or internal links to the corresponding pages.

## References

1. Meusel, R., Petrovski, P., Bizer, C.: The WebDataCommons Microdata, RDFa and Microformat Dataset Series. In: Proceedings of the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Trentino, Italy, Springer Berlin Heidelberg (2014) 277–292
2. Mika, P., Potter, T.: Metadata Statistics for a Large Web Corpus. In Heath, T., Bizer, C., Berners-Lee, T., eds.: Proceedings of the WWW2012 Workshop on Linked Data on the Web (LDOW 2012), Lyon, France (2012)
3. Goel, K., Guha, R.V., Hansson, O.: Introducing Rich Snippets. http://googlewebmastercentral.blogspot.de/2009/05/introducing-rich-snippets.html (2009)
4. 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), Beijing, China, ACM (2011) 725–734
5. Meusel, R., Vigna, S., Lehmberg, O., Bizer, C.: Graph Structure in the Web – Revisited: A Trick of the Heavy Tail. In: Proceedings of 23rd International World Wide Web Conference (WWW 2014 Companion Volume), Seoul, Korea, ACM (2014) 427–432
6. Buck, C., Heafield, K., van Ooyen, B.: N-Gram Counts and Language Models from the Common Crawl. In: Proceedings of 9th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavk, Iceland (2014) 26–31
7. Common Crawl: Frequently Asked Questions. http://commoncrawl.org/faqs/

8. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. techreport, Stanford InfoLab (1998)
9. Harth, A., Umbrich, J., Decker, S.: MultiCrawler: A Pipelined Architecture for Crawling and Indexing Semantic Web Data. In: Proceedings of the 5th International Semantic Web Conference (ISWC 2006), Athens, GA, USA, Springer Berlin Heidelberg (2006) 258–271
10. Meusel, R., Mika, P., Blanco, R.: Focused Crawling for Structured Data. In: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM 2014), Shanghai, China, ACM (2014) 1039–1048
11. Dalvi, N., Machanavajjhala, A., Pang, B.: An Anaylsis of Structured Data on the Web. Proceedings of the VLDB Endowment **5**(7) (2012) 680–691
12. Baeza-Yates, R., Castillo, C.: Crawling the Infinite Web: Five Levels Are Enough. In: Proceedings of the 3rd International Workshop on Algorithms and Models for the Web-Graph (WAW 2004), Rome, Italy, Springer Berlin Heidelberg (2004) 156–167
13. Web Data Commons: Web Data Commons - RDFa, Microdata, and Microformats Data Sets - December 2014. http://www.webdatacommons.org/structureddata/2014-12/stats/stats.html
14. Berners-Lee, T., Fielding, R.T., Masinter, L.: Uniform Resource Identifier (URI): Generic Syntax. http://www.ietf.org/rfc/rfc3986.txt (2005)