Understanding the Impact of E-Commerce Software on the Adoption of Structured Data on the Web

Kurt Uwe Stoll, Mouzhi Ge, and Martin Hepp

Universität der Bundeswehr München
E-Business & Web Science Research Group
Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany
{uwe.stoll|mouzhi.ge}@unibw.de, mhepp@computer.org

Abstract. In this paper, we analyze the potential impact of e-commerce software packages on the diffusion of markup for structured data on the Web. We argue that such an analysis must focus on the *product detail pages,* i.e. the "deep links" to individual items, rather than the pure number of shop sites, for assessing the potential. Based on (1) a systematic analysis of the popularity of 56 software packages for e-commerce sites among the Alexa list of the one million most popular Web sites, we (2) estimate the number of product detail pages for the respective shops and then (3) project the potential lever of each e-commerce package for the adoption of structured markup, assumed that adding support for structured data can be made a readily available feature of a respective software. Our results indicate that by adding a structured markup component to as little as seven popular e-commerce systems, we could instantly deploy structured data markup on nearly 90 % of the product detail pages among the one million most popular Web sites.

Keywords. E-Commerce, Semantic Web, RDFa, Microdata, Microformats, SEO, schema.org, GoodRelations, Technology Adoption

1 Introduction

In the past few years, embedding structured data with e-commerce information into HTML content using RDFa[1], Microdata[2], or Microformats[3] has become a mainstream technique for Web shops. With GoodRelations [1], an established ontology is available that supports modeling of a wide range of e-commerce scenarios with structured data. It has recently been integrated into the schema.org standard [2] advocated by four major search engines.

For shop owners, publishing structured data is mainly motivated by the consumption of search engines like Google, which use it to enhance search results. From a search engine perspective, shop pages that contain structured data are preferable, as the extraction of important information like product name or price is computationally expensive and error-prone. When pages include structured data markup, this task becomes less difficult and potentially more reliable.

[1] http://www.w3.org/TR/xhtml-rdfa-primer
[2] http://www.w3.org/TR/microdata/
[3] http://microformats.org/about

In the context of the paper, we define e-commerce software packages as software artifacts that allow merchants to operate an e-commerce site that presents products or services and supports a purchasing transaction, e.g. via shopping cart functionality. Many shop operators use *standard* e-commerce applications to run their sites, such as Magento[4], ATG[5] or Prestashop[6]. State-of-the-art solutions typically offer an extension mechanism that enables adding third-party code to the resulting system. This allows developing extension modules for structured data, which then significantly reduce the cost and effort of adding structured data markup, because instead of editing the HTML template, the administrator will just have to download and install the respective module. In the past years, we have developed or helped others to develop several such modules, which have shown significant uptake. As of January 2013, there are more than 18,000 aggregated downloads for our modules for different e-commerce applications that add structured markup. Still, these 18,000 downloads represent only a small part of the total number of shop sites on the Web.

For advancing the field of Semantic Web technology in e-commerce, however, the adoption of structured data markup by a limited number of Web shops is not enough, as the resulting market coverage will not be sufficient in terms of the range of products, or the coverage of dealers with different value propositions. A broad range of products is especially relevant for applications that aim at providing a consolidated view on a respective market segment. For instance, a product search engine based on structured data for digital cameras will only be useful if a substantial share of the market is represented.

Our core research question is how e-commerce packages and the respective market structures provide an effective lever for accelerating the diffusion of structured data markup so that Semantic Web applications become possible. In essence, we want to know the number of e-commerce sites that run standardized software packages, and the distribution properties of the number of products per shop site. This allows projecting the impact of adding structured data functionality to a comparatively small number of codebases of e-commerce packages. To our knowledge, no previous systematic analysis of these questions exists, except for non-scientific studies from an industry perspective.

In this paper, we provide (1) a systematic analysis of the popularity of 56 e-commerce applications among the Alexa list of the one million most popular Web sites [3], (2) estimate the number of product detail pages for the respective shops and then (3) project the potential lever of each package for the adoption of structured data markup, assumed that adding support for structured data can be made a built-in feature of a respective application. Our results show that by adding a structured markup component to as little as *seven* popular Web shop applications could instantly add structured data to nearly 90 % of the product detail pages among the one million most popular Web sites. Our main contribution is estimating the impact of adding structured data functionality to a limited number of e-commerce packages for the amount

[4] http://www.magentocommerce.com

[5] http://www.oracle.com/us/products/applications/commerce/atg/

[6] http://www.prestashop.com/

of respective markup at Web scale, which resides mainly in the "deep link" part of Web sites.

The remainder of this paper is structured as follows: In section 2, we summarize related work. In section 3, we describe our methodology and provide details about our data collection and implementation. In section 4, we analyze the data and summarize the findings. In section 5, we evaluate the performance of the critical component for the detection of e-commerce systems, discuss potential limitations of our approach, and sketch future work. In section 6, we conclude and summarize our results.

2 Related Work

In this section, we summarize work related to our approach, which can be grouped into three categories:

Market Studies: Due to the very dynamic nature of the field, it is unfortunately inevitable to refer to non-scientific resources for some figures. For instance, Raju estimates that in 2012, there were 90.500 shops in the US earning more than $12,000 [4].

Since 2011, Robertshaw has been conducting a semi-annual analysis of the market shares of e-commerce systems [5]. According to his results, the eleven biggest e-commerce systems account for more than 80% of all sites. The service http://builtwith.com provides ongoing reports of the popularity for a wide range of Web technologies, including e-commerce packages [6]. Unfortunately, the site only delivers relative market share data of the ten most popular e-commerce packages with respect to the top 1 Million sites sample, which is of limited value for our research. Note that all these reports do not take into account the size of the "deep link" part of shop sites but merely count the sites directly.

Functional Comparison: Beside the market studies, there are many analyst publications targeting e-commerce systems, mostly aimed at corporate audiences. Those publications put a stronger focus on the comparison of features and a strategic outlook on the regarded systems than e.g. on the number of deployments. For instance, in 2011 Gartner [7] provided a report that maps different e-commerce systems into four clusters. Since the criteria for inclusion in the report are such that they exclude lower end solutions, which may account for a substantial amount of Web shops in the long tail, the study does not match our focus. In 2012, Forrester [8] provided a similar report ranking different solutions in terms of offering and strategic position, also excluding lower end solutions.

Structured Data for E-Commerce: The GoodRelations Web vocabulary [1] was the first broadly adopted Web ontology for exposing structured data with e-commerce information following the Semantic Web vision [9][10]. The adoption of GoodRelations was supported by a wealth of freely available tools for publishing and consuming respective data; for an overview, see [11]. While so far there is no comprehensive quantitative analysis of the amount and granularity of GoodRelations data in the wild, Ashraf et al. have provided a preliminary study on GoodReations usage patterns on the Web [12]. Recently, GoodRelations has been officially integrated into schema.org [2]. Schema.org is a collaboration of Google, Yahoo!, Bing and Yandex to promote a set of stable structured data vocabularies. We estimate that as of this writing, at least

15,000 shop sites with a total of at least 20 million product detail pages include GoodRelations data.
To the best of our knowledge, there is no previous work that analyzes the impact of e-commerce systems on the availability of structured data.

3 Methodology, Approach, and Implementation

The basic rationale for our research is the following: We know from our previous development of extension modules that it is possible to modify respective shop software packages to automatically add the publication of structured data based on GoodRelations, and that in a manner that (1) requires only minimal configuration effort for the site owner and (2) is tolerant with regard to modifications of the stylesheets, themes, and HTML templates, or the installation of other modules. Our next goal is to add respective functionality to the core codebase of popular e-commerce packages so that the adoption of structured data does no longer depend on the manual installation of such extensions. If we succeed with that, a large number of shop sites will automatically add GoodRelations markup once they are updated to the next version of the system. This promises to be a huge lever for the implementation of the Semantic Web vision for e-commerce. Given that we have limited resources for implementing the idea, we need to know (1) how many e-commerce systems we should target and (2) which coverage we can achieve on the level of product detail pages. We expect the number of those pages to be Pareto-distributed, likewise than, e.g. firm sizes. Thus, covering a minor share of systems with structured data may have a high impact on the overall coverage of the market.
There are four layers of interest in this context, as Fig. 1 illustrates.

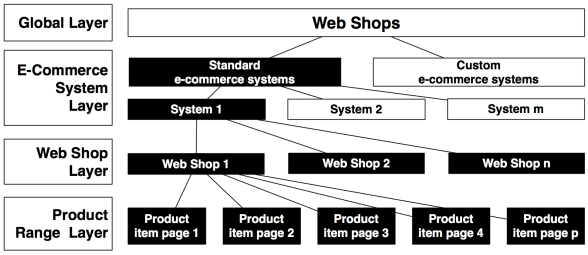

Fig. 1. Effect of enabling structured data for an e-commerce system on product pages.

1. The global layer, which represents all shops on the Web.
2. The e-commerce system layer, which is divided into Web shops running standard e-commerce packages, and custom e-commerce systems based on proprietary software. In our work, we focus on standardized e-commerce systems, such as Magento, ATG, or Prestashop.
3. The third layer is the Web shop layer, constituted by the actual shop sites that run a specific e-commerce package or proprietary implementation.
4. The fourth layer is the product range layer. It consists of all the product detail pages hosted by a particular shop site and system.

3.1 Methodology

Our research approach consists of the following steps:

1. **Obtaining a list of relevant site URIs:** Since we cannot analyze the Web as a whole, we need a subset of URIs representing Web site main pages to start with. Roughly speaking, this is a list of Web sites, but not limited to e-commerce sites. We will screen them for e-commerce functionality in the subsequent steps. For our analysis, we take the freely provided Alexa Top 1 million traffic rank [18]. This gives us the URIs of the main pages of the one million most popular sites.
2. **Defining the shop software packages to search for:** As a second step, we need a list of relevant e-commerce software applications. For that purpose, we merged the top 40 list provided by Robertshaw [16], the top ten list from http://www.builtwith.com and the systems mentioned in the reports by Gartner [2] and Forrester [3], resulting in a list of 56 search strings for e-commerce systems. This list is shown in Table 1.
3. **Determining whether a URI represents a Web shop using one of the systems from our list:** To get a hold of URIs that are run by specific e-commerce systems, we used the tool Whatweb [19], originally a site profile scanner from the context of computer security. Whatweb is able to detect a wide range of properties of a Web-site, including e-commerce functionality. We then matched the results against the list of search strings.
4. **Counting product item pages based on sitemaps:** Next, we need to estimate the number of product detail pages for each shop, which is a non-trivial challenge. As an approximation, we used XML sitemaps [20] of the shop sites, if available. In this context, we assumed the remaining sites to be a sufficient sample of the base population. We then conducted a cluster analysis on the sitemap properties to find the ratio of product item pages on a Web shop. We could show that product item pages and overall sitemap pages correlate. Thus, we use the URI counts based on sitemap in combination with the average share of product detail pages within a site as a first approximation of the number of products per shop.
5. **Extrapolation of the product item count to Web scale:** In order to predict the impact of equipping e-commerce systems with structured data markup, we project our results on the total number of shops in the population.
6. **Evaluation of the e-commerce system detection:** As the e-commerce system detection provided by Whatweb is a critical part of our analysis, we additionally evaluate its performance on a sample of n=550 URIs with human computation.

Table 1. Consolidated list of search strings for the 56 e-commere packages in the study.

magento, zen cart, virtuemart, oscommerce, prestashop, opencart, volusion, Yahoo! stores, interspire, ubercart, wp e-commerce, ecshop, actinic, miva merchant, shopify, cs-cart, ibm websphere commerce, x-cart, oxid esales, 3dcart, atg, demandware, ejunkie, intershop, shopp, ablecommerce, nopcommerce, pro-stores, shopsite, foxycart, big cartel, ekmpowershop, gsi commerce, shopfactory, cubecart, romancart, tomatocart, drupal commerce, blucommerce, lemonstand, thefind upfront, google trusted store, clever-bridge, elastic path, icongo, jagged peak, marketlive, microsoft commerce server, netsuite, istore, venda, micros-retail, redprairie, digital river, sap e-commerce, xt-commerce

3.2 Implementation

Obtaining a list of relevant site URIs: The initial input to our study is the top one million list of Alexa [13]. Alexa analyzes website popularity. There is a monthly global ranking of top-level domains according to traffic (the "Top1m list"), provided for free in a CSV format. We used the 09/2012 release.

To understand which e-commerce packages are used for the Web shops in the Top1m list, we employed the tool Whatweb [14]. Whatweb is an open-source security scanner written in Ruby. Among other site characteristics, it detects server software, content management systems, and e-commerce systems.

Unfortunately, applying a tool like Whatweb on such a large amount of URIs is computationally expensive. Thus, we employed cloud computing resources. While a common pattern is to distribute the task on many cloud instances, we found that running it parallelized on a single powerful machine was sufficient and resulted in the smallest overhead. We used the Amazon EC2 Cluster Compute Eight Extra Large cloud computing instance (cc2.8xlarge)[7]. We distributed four threads on each of the 16 cores of the machine using GNU parallel[8] with one line of code, which can be found in listing 1. Running the task took 8 hours and 32 minutes, resulting in server cost of 19.20 $ for the given 1 million URIs.

```
cat 1m.csv | parallels --max-threads=64 ruby whatWeb.rb > 1m.txt
```

Listing 1. Parallelization with GNU parallel.

To get the subset of results related to the e-commerce packages of interest, we merged the top 40 list provided by Robertshaw, the top ten of builtwith.com and the leading systems from the Gartner and Forrester reports, as already mentioned. The merged list of 56 search strings is given in Table 1 above.

For consistency, in the tables and figures of the remainder of the paper, we use the original lower case spellings of the search strings. We additionally considered the system *XT commerce*, as it was missing in the other surveys and is claimed to have more than 100,000 installations. We ran the list against the Whatweb results using a small script, matching the search strings against the Whatweb result file. It is important to stress that, to a certain degree, this approach is also able to detect shop systems even if there was no specific Whatweb plugin beforehand, as there are often strings hinting to shop systems in the part of the results of Whatweb (eg. cookies or HTTP headers). Those were not targeted by the original server detection plugins.

Counting product item pages based on XML sitemaps: After fetching and parsing the sitemaps, we went on to get a hold of product item pages. This figure is important, as web shops usually provide, beside the product detail pages (1) category pages, (2) review pages and (3) pages about payment and shipping options, to name a few. In order to assess the number of product detail pages in a given Web shop we assumed that the product item pages count should be correlated to the total URI count of the XML sitemap. To validate this, we conducted a k-means cluster analysis on the prop-

[7] http://aws.amazon.com/en/ec2/#instance
[8] http://www.gnu.org/software/parallel/

erties of each entry of the sitemap of a sample of 716[9] randomly selected shops using Scikit-learn 2011 [15]. We set the cluster size to three, as we assumed there would be a cluster of product pages, category pages, and of arbitrary pages. In preprocessing, we filtered URIs linking to images, and generated a property that yielded the existence of the string "product" in the server path. For the further analysis, we only took product, priority and lastmod properties into account. The resulting clusters were filtered to have a silhouette coefficient [16] of at least 0.6, and the relative size of the biggest cluster to be between 0.6 and 0.9 of the number of entries in the sitemap, as we considered only those who match this threshold as valuable sitemaps matching our initial assumptions. We then computed Pearson's correlation between the biggest cluster and the total sitemap page count. This resulted in a value of 0.879, indicating a strong correlation. Additionally, we computed a final correction factor that represents the mean difference between URIs found in a sitemap and its biggest cluster. The result is 0.774, with a 95% confidence interval of 0.759 to 0.790. Thus, in 95 % of the caes, there will be between 759 and 790 product item pages per 1000 URIs in a XML sitemap.

4 Results

4.1 Summary

Number of products per site: According to the correlation analysis conducted in section 3.2, we can take the URIs listed in a XML sitemap as an estimate for the number of product detail pages. The analysis of the XML sitemaps gives preliminary hints that the market for e-commerce systems follows a Pareto distribution with regard to the number of product detail pages, i.e. at the level of deep links. Six systems leading the URI count represent more than 90% of all URIs. The respective results are shown in table 2. Overall, 23.33 million URIs could be extracted from the XML sitemaps. If we apply the correction factor of 0.774 (see section 3.2), this projects to roughly 18 million product item pages.

Table 2. URIs found in sitemaps and product item estimate

	Shop software	**URIs**	Lower boundary of 95% confidence interval	**Projected # of product item pages (n * 0.774)**	Upper boundary of 95% confidence interval	**% of products of all products**	**Cumulated % of products**
1	Magento	12,610,254	9,571,183	9,760,336	9,962,101	54.05	54.05
2	ATG	3,016,552	2,289,563	2,334,811	2,383,076	12.93	66.98
3	Prestashop	2,756,334	2,092,058	2,133,402	2,177,504	11.81	78.80
4	osCommerce	1,597,558	1,212,547	1,236,509	1,262,071	6.85	85.64
5	Zen Cart	769,947	584,390	595,938	608,258	3.30	88.94
6	CS-Cart	524,778	398,307	406,178	414,575	2.25	**91.19**
7	Virtuemart	508,310	385,807	393,431	401,565	2.18	93.37
8	Others	1,546,366	1,173,692	1,196,887	1,221,629	6.63	100
Total		**23,330,099**	**17,707,545**	**18,057,496**	**18,430,778**		

Additionally, we visualized the findings using box plots [17], as shown in Fig. 2. For a higher expressiveness of the plot, we filter the systems to have product page counts in the 0.5 area of the standard deviation of each system's distribution, i.e. we filtered out extremely large (and small) sites. The resulting set of eight systems is the result of

[9] This number emerged from n=50 samples per shop maximum, if available.

applying a filter so that only shops that have more than 50 results are considered. The boxes show the 50% quantile of the distributions after applying the filter above, the line in the box the median. The lines above and below the boxes are the whiskers, they show the remaining upper and lower 25% quantiles. Outliers are plotted as crosses. We can see that Demandware aggregates a high amount of URIs and additionally has high top 25% whiskers, whereas Virtuemart or Zen Cart do not. As the median of the distributions is mostly located considerably towards the bottom of the 50% box, those systems spot a positive skew towards a low number of products. It also matches our informal experiences with maintaining various shop extensions.

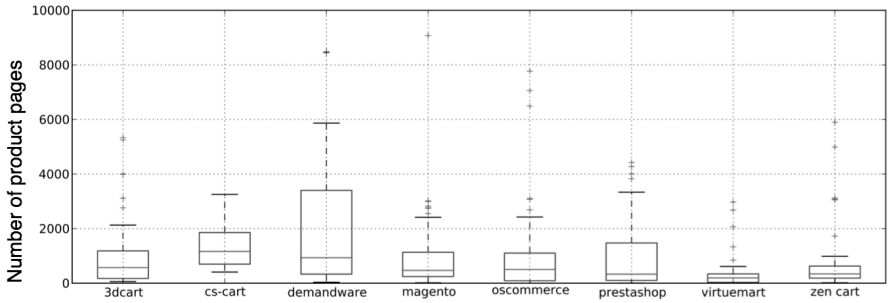

Fig. 2. Distribution of the number of product pages per shop software package

Market structures and popularity: The initial Whatweb experiment resulted in 912,865 successful responses. Overall, 21,848 shops could be detected in the sample. The frequency count of the different e-commerce systems is shown in table 3. Only six e-commerce systems cover more than 80% of regarded sample, and Magento leads the results with 34.7%.

Table 3. Popularity of e-commerce software applications

System	Magento	osCommerce	Prestashop	Virtuemart	Zen Cart	XT commerce	Opencart	CS-Cart	Others	Sum URIs
URIs	7,582	2,764	2,629	2,144	2,129	1,131	545	455	2469	21,848
%	34.70	12.65	12.03	9.81	9.74	5.18	2.49	2.08	11.30	
sum %	34.70	47.35	59.39	69.20	78.95	**84.12**	86.62	86.62	88.70	

4.2 Impact of E-commerce Software on the Adoption of Structured Data

Based on the tentative findings regarding the number of product detail pages and market structure, we can assume that adding structured data to the core codebases of only six e-commerce systems would already augment nearly 90 % of the product detail pages found in the sample with structured data markup. The systems with the highest impact would be Magento, osCommerce, ATG, Zen Cart, Prestashop, EC-Shop and Virtuemart. Except for ATG and EC-Shop, for all of these there are already GoodRelations extension modules available [18]. However, only a small share of shops actually use those extensions. Therefore, it should be a priority to add structured data functionality to the core codebase of shop systems. Another important activity will be the development of missing extensions.

4.3 Additional Findings

Site popularity: We analyzed the popularity of sites generated by specific e-commerce systems in terms of the Alexa traffic ranking. Herein, it is of interest which shop systems tend to be more present in the high-traffic sites, and which not. To answer this question, we chose to use the mean of each shops ranking distribution (AX-mean). To make the result more transparent, we provide an additional variable AX-factor, defined by dividing 500,000, i.e. the middle rank of the Alexa traffic ranking, with the mean of each shop. A higher value means higher ranking sites in average. Most shops ranked less than 1, which means that most of them position in the lower ranks of Alexa Top1m sites. Only ATG and Demandware yielded significant values of 2,7227 and 1,634, indicating that they are used by highly popular shops. A possible explanation is that really large shop applications use either proprietary code or employ technology components like load balancers that render the underlying e-commerce system hard to detect.

Sitemap availability and quality: Another observation is that many sites yielded incorrect sitemaps or provided none at all. To gain insight into this, we tried to fetch the sitemap according to robots.txt or the standard "sitemap.xml" path [19]. Then, we analyzed the results and counted occurrences of URIs. We additionally implemented an algorithm to parse sitemap indices. Only the previously low ranked 3DCart yielded significant positive results. The other shops achieved rates lower than 67%, down to, most often 50%. This means the sitemap standard is not used properly in many cases, and results in high crawling effort for search engines. We expected high-end systems such as ATG to provide correct sitemaps, but their results were not better than those of the base sample.

Geographical distribution: Finally, we analyzed the geographical distribution of leading systems according to the number of products they manage. The United States, Germany, United Kingdom and France dominate the geographical distribution. The top ten countries of the seven most popular e-commerce systems are shown in Table 5.

Table 5. Geographical distribution of the systems according to the number of products

	Magento		**ATG**		**Prestashop**		**osCommerce**		**Zen Cart**		**CS-Cart**		**Virtuemart**	
1	USA	**3062**	USA	**456**	FRAU	**950**	USA	**946**	USA	**1180**	USA	**226**	USA	**615**
2	GER	**974**	GER	**266**	USA	**436**	GER	**427**	UK	**116**	UK	**47**	GER	**305**
3	UK	**693**	UK	**44**	GER	**265**	FRA	**268**	NL	**110**	GER	**34**	RUS	**219**
4	FRA	**643**	FRA	**23**	SPA	**147**	UK	**239**	GER	**77**	AUS	**20**	FRA	**146**
5	NL	**371**	RUS	**19**	UK	**88**	POL	**99**	EST	**57**	NL	**15**	UK	**107**
6	BRA	**158**	CAN	**18**	POL	**68**	SPA	**90**	ROC	**49**	VN	**14**	NL	**65**
7	EU	**155**	TUR	**18**	EU	**52**	NL	**86**	EU	**38**	SA	**11**	UKR	**49**
8	AUS	**138**	NL	**17**	CZ	**48**	EU	**43**	MAL	**32**	EU	**9**	HU	**47**
9	SPA	**119**	EU	**16**	NL	**48**	AUS	**41**	JAP	**29**	POR	**8**	IT	**45**
10	IRE	**77**	POL	**12**	CAN	**42**	RUS	**41**	CZ	**28**	GR	**5**	POL	**45**

5 Evaluation of Shop Software Recognition

The e-commerce system detection is a critical part of our approach. We decided to assess the performance of the method using human computation via the service

Crowdflower[10], which is an intermediary providing access to a manifold of human computation services through a standardized interface. We use precision to evaluate the performance of information retrieval systems [20], as we cannot measure recall, because our approach is limited by the aforementioned list of systems. We set up a task for humans to decide whether a given URI is an e-commerce site or not. Thus, the experiment provides insight whether the list of shop URIs actually contains shop sites.

We ran the experiment for 11 e-commerce systems and presented to the human participants a list of 50 randomly selected URIs, resulting in 550 items to judge. According to the evaluation, the shop detection approach achieved a mean precision of 94%, i.e. the shops detected by Whatweb are actually shops. The systems we analyzed yielded a precision between 90% (e.g. ATG, Virtuemart) and 100% (e.g. 3DCart and Demandware). We show the results in Table 6.

Table 6. Reliability of the shop detection technique

Shop Software	3DCart	Demandware	Shopsite	Magento	Prestashop	CS-Cart	
Precision	100.00 %	100.00 %	97.00 %	96.00 %	93.00 %	92.00 %	
Shop Software	EC-Shop	osCommerce	Zen Cart	ATG	Virtuemart		**Mean**
Precision	92.00 %	92.00 %	92.00 %	90.00 %	90.00 %		94.00 %

6 Discussion and Conclusion

6.1 Limitations

Our work is subject to the following limitations:

1. Alexa Top1m as basis for the data collection induces is a bias towards popular sites. As future work, we plan to run Whatweb against the data of CommonCrawl [21], a public crawl of a substantial part of the Web. This would lower the bias towards popular sites and better represent the long tail of the Web.
2. We used Whatweb as it is, without additions to the plugins or constraining functionality. Improving the plugins could have resulted in higher performance in the site recognition process, but the overall result of our research is not dependent on marginal performance improvements of the underlying data collection. Constraining functionality of Whatweb in terms of excluding detection features would have resulted in a lower computational effort, but we would have lost additional data, which can be explored in future work.
3. E-commerce software missing in our initial links, additional components like load balancing tools, or weaknesses in the recall of our detection technique may account for a significant number of sites incorrectly excluded from our analysis. We evaluated this by drawing a sample of 100 sites of the Alexa list and manually judged whether they were shop sites. This resulted in a share of 21.74% e-commerce sites as compared to only 2.39% sites found by our technique. This may reflect a fundamental limitation of our quantitative results, unless the shop

[10] http://www.crowdflower.com

sites properly detected are a sufficiently representative sample of the overall situation. At this point, we do not know this.

4. Another shortcoming might be the reliability of the string search over the results of Whatweb in order to detect the different shop systems.
5. The approach of using XML sitemaps to estimate the number of deep product detail pages is a limited technique. Many sites to not provide XML sitemaps and XML sites provided may list only a subset of actually available product item URIs. Alternative approaches for counting the number of product detail pages would be (1) deep crawling or (2) counting pages indexed by Google[11]. In order to evaluate the quality of our approach, we conducted a deep crawl and requested the pages indexed by Google of n=13 randomly selected sites generated by five different shop packages. We then computed the correlation coefficients between the sitemap count and the two new heuristics. The results are shown in Table 7.

Table 7. Correlation of differerent page count heuristics

	Sitemap count / Deep crawl	Sitemap count / Google	Deep crawl / Google
Pearson's correlation	0.38	0.72	0.45

We see that neither (1) a sitemap count vs. a deep crawl nor (2) a deep crawl vs. Google index size correlate significantly. Only sitemap count vs. Google index size shows a more significant correlation of 0.72. To clarify those results, we manually inspected the URIs in the deep crawl. However, a deep crawl count without reliably detecting product detail pages seems to be a very questionable heuristic, as shops often (a) generate pages for every permutation of a category filter, and (b) provide pages for every review submitted by customers. This leads to giant URI lists from deep crawls stemming from duplicate content. Also, the number of pages indexed by Google is a problematic estimate, as it depends on the index size and crawling budget Google allocates to a site. We further assessed the objection that the page count might differ between (1) sites that provide sitemaps and (2) sites that do not. An additional deep crawl on n=10 sites resulted in means of (1) 1,620.69 and (2) 2,110.20 and standard deviations of (1) 1,854.39 and (2) 4,543.26. This finding does not harm our initial result. To summarize, we think the approach taken is a fair technique given the limitations of alternative solutions.

6.2 Conclusion

In this paper, we presented an extensive analysis of Web shops within the one million most popular sites, in order to assess the impact of standardized e-commerce systems on the adoption of structured data markup for the Semantic Web. We can show that based on the high number of product detail pages, i.e. the "deep links" in shopping sites, adding structured data functionality to only six e-commerce software packages could add structured data markup to more than 90 % of all products detail pages in the sample. We have shown that working on the integration of the Semantic Web vision into those six software packages will likely be a very effective lever for the diffusion

[11] We requested the figure with Google site search, i.e. +site:example.org

of structured data markup into real applications and to increase the market coverage in the resulting data.

Acknowledgments: The work on this paper has been supported by the German Federal Ministry of Research (BMBF) by a grant under the KMU Innovativ program as part of the Intelligent Match project (FKZ 01IS10022B), and by the Eurostars program (within the EU 7th Framework Program) of the European Commission in the context of the Ontology-based Product Data Management (OPDM) project (FKZ 01QE1113D).

References

1. Hepp, M.: GoodRelations: An Ontology For Describing Products And Services Offers On The Web. In: Knowledge Engineering: Practice and Patterns. Vol. 5268. pp. 329-346. Springer Berlin Heidelberg. (2008).
2. http://schema.org/, last checked on 01/10/2013.
3. Alexa Top 1 Million Sites By Traffic Rank As CSV, last checked on 01/10/2013.
4. How Many Online Stores Are There In The U.S.? http://blog.referralcandy.com/2012/08/14/how-many-online-stores-are-there-in-the-u-s/, last checked on 01/10/2013.
5. Tom Robershaw: October 2012 Ecommerce Survey, http://tomrobertshaw.net/2012/11/october-2012-ecommerce-survey/, last checked on 01/10/2013.
6. Ecommerce Technology Web Usage Statistics, http://trends.builtwith.com/shop, last checked on 01/10/2013.
7. Alvarez, G., Fletcher, C., Sengar, P., Martz, S. A.: Magic Quadrant For E-Commerce. Gartner, Inc. (2011).
8. Walker, B. K.: The Forrester Wave (™): B2C Commerce Suites, Q3 2012. Forrester Research (2012).
9. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American, 284. 5. pp. 34-43 (2001).
10. Lassila, O.: Web Metadata: A Matter Of Semantics. IEEE Internet Computing, 2. 4. pp. 30-37 (1998).
11. Hepp, M., Radinger, A., Wechselberger, A., Stolz, A., Bingel, D., Irmscher, T., Mattern, M., Ostheim, T.: GoodRelations Tools And Applications. Poster and Demo Proceedings of the 8th International Semantic Web Conference (ISWC 2009), Washington, DC, USA, (2009).
12. Ashraf, J., Cyganiak, R., O'riain, S., Hadzic, M.: Open Ebusiness Ontology Usage: Investigating Community Implementation Of GoodRelations. In: Proceedings of the WWW2011 Workshop on Linked Data on the Web (LDOW2011). Vol. 813 (2011).
13. Can I Get A List Of Top Sites Using Web Services?, http://www.alexa.com/faqs/?p=35, last checked on 01/10/2013.
14. Whatweb, http://www.morningstarsecurity.com/research/whatweb, last checked on 01/10/2013.
15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-Learn: Machine Learning In Python. Journal of Machine Learning Research, 12. pp. 2825-2830 (2011).
16. Rousseeuw, P.: Silhouettes: A Graphical Aid To The Interpretation And Validation Of Cluster Analysis. J. Comput. Appl. Math., 20. 1. pp. 53-65 (1987).
17. McGill, R., Tukey, J. W., Larsen, W. A.: Variations Of Box Plots. The American Statistician, Vol. 32. Issue 1. pp. 12-16 (1978).
18. Adding GoodRelations To Standard Shop Software, http://wiki.goodrelations-vocabulary.org/Shop_extensions, last checked on 01/10/2013.
19. Sitemaps.Org - Protocol, http://www.sitemaps.org/protocol.html, last checked on 01/10/2013.
20. Manning, C. D., Raghavan, P., Schütze, H.: Introduction To Information Retrieval. Cambridge University Press (2008).
21. CommonCrawl, http://commoncrawl.org/, last checked on 01/10/2013.