A Web-scale Study of the Adoption and Evolution of the schema.org Vocabulary over Time

Robert Meusel
Data and Web Science Group
University of Mannheim
B6 26, Mannheim, Germany
robert@informatik.uni-mannheim.de

Christian Bizer
Data and Web Science Group
University of Mannheim
B6 26, Mannheim, Germany
chris@informatik.uni-mannheim.de

Heiko Paulheim
Data and Web Science Group
University of Mannheim
B6 26, Mannheim, Germany
heiko@informatik.uni-mannheim.de

ABSTRACT

Promoted by major search engines, schema.org has become a widely adopted standard for marking up structured data in HTML web pages. In this paper, we use a series of large-scale Web crawls to analyze the evolution and adoption of schema.org over time. The availability of data from different points in time for both the schema and the websites deploying data allows for a new kind of empirical analysis of standards adoption, which has not been possible before. To conduct our analysis, we compare different versions of the schema.org vocabulary to the data that was deployed on hundreds of thousands of Web pages at different points in time. We measure both top-down adoption (i.e., the extent to which changes in the schema are adopted by data providers) as well as bottom-up evolution (i.e., the extent to which the actually deployed data drives changes in the schema). Our empirical analysis shows that both processes can be observed.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Systems—*Web-based Services*; C.2.6 [**Computer-Communication Networks**]: Internetworking—*Standards*

General Terms

Standardization, Measurement

Keywords

Microdata, schema.org, Standardization, Adoption, Data Space Profiling

1. INTRODUCTION

In the recent years, languages for incorporating structured knowledge into HTML web pages, such as RDFa, Microformats, and Microdata, have been proposed. Out of those, the latter shows the widest adoption [15], in particular due to the schema.org initiative driven by major web search engines such as Google, Bing, Yahoo!, and Yandex.[1] schema.org defines a common set of classes and properties to mark up web mostly business related contents, such as companies, addresses and opening hours, or product offers.

The *Common Crawl Foundation*[2] issues publicly available, large-scale web crawls covering billions of pages. From those crawls, the Web Data Commons project regularly extracts structured data, such as Microdata, Microformats, and RDFa.[3]

Since both the schema as well as the deployed data on the Web are publicly available at different points in time, this allows for *a new methodology of analyzing standard adoption*: instead of sending questionnaires to possible adopters and analyzing the responses (where the number of responses are usually small, in particular for longer and deeper questionnaires), we can observe the adoption directly from the data, published by hundreds of thousands of standard adopters.

The main motivation for web site providers to include Microdata is an improved displaying of results by major search engines and by this an improved awareness of their page to the user. Search engines display richer results for web sites described with Microdata, i.e., those web sites are presented more prominently to the user [10].

Since its start in 2011, schema.org has undergone more than 25 revisions, ranging from small typo fixes in schema elements to the integration of entire new vocabularies for specific domains, such as the Music Ontology[4] or the GoodRelations vocabulary[5]. Besides adding new elements, the usage conventions of existing elements are sometimes change to fit the rest of the standard better. Additionally, elements whose use is no longer encouraged are occasionally marked as deprecated or, more often, as superseded by others.

At the same time, web data providers use schema.org to mark up data on the Web. As shown in [14], the actually deployed data can heavily deviate from the standard definitions. Frequent deviations include the usage of undefined classes and properties, as well as the usage of elements in a context in which they are not supposed to be used.

In this paper, we add a *diachronic perspective* to have a closer look at those phenomena. Using snapshots of the

Web Intelligence, Mining, and Semantics (WIMS) Cyprus, 2015

[1] http://schema.org
[2] http://commoncrawl.org/
[3] http://webdatacommons.org/
[4] http://musicontology.com/
[5] http://www.heppnetz.de/projects/goodrelations/

schema and deployed data from different points in time, we analyze both how new elements in the standard are adopted, and how the use of undefined elements influences the evolution of the standard.

The schema definition is maintained in a community-driven process, with prospective changes being announced and discussed in public mailing lists. However, the number of stakeholders taking part in those discussions is considerably small given the thousands of data providers *using* the schema.

In this paper, we attempt an empirical, *data-driven* analysis of the interaction between those two groups. More specifically, we look at *top-down* and *bottom-up* processes. For the former, we analyze how fast and to which extent changes in the schema are adopted by data providers. For the latter, we examine how strongly changes in the schema are driven by undefined, yet frequent usages of schema elements. Wherever possible, we also try to find influences driven by the *data consumers* – e.g., the tutorials provided by search engine providers such as Google. Making use of this novel two-sided methodology to analyze the adoption and evolution of the deployment of schema.org over last four years we reveal useful insights about the data space.

Those insights on the one hand side, help to understand the state-of-the art deployment. On the other hand side, it allows the prognosis for the further development in this area.

From a broader perspective, we show that the availability of deployed data on the web allows for a new kind of data-driven, empirical analysis of standard adoption and evolution, which reveals interactions between the standardization and the actual usage of the standard.

The rest of this paper is structured as follows. In Section 2, we describe the data corpus used for the analysis in this paper. Section 3 defines the research questions to be analyzed and the measures we use, and in Section 4, we present and discuss our empirical findings. We review related work in Section 5, and conclude with a summary and an outlook on future research.

2. DATA CORPUS

In order to analyze the evolution of schema.org Microdata, we make use of the Microdata corpora extracted by the *WebDataCommons* project from the public available Web crawls of the *Common Crawl Foundation*. Thus, we base our data on deployments on a corpus of billions of web sites, among which several hundreds of thousands deploy the schema.org vocabulary.

The datasets we use include all information marked up directly within the HTML page using the Microdata markup language[6]. From those pages, structured data is extracted using the *Apache Anything to Triples (Any23)* library[7]. The data is stored using the RDF N-Quads format[8], which combines the `subject`, `predicate` and `object` of a standard triple with the `graph` information, which for those datasets is the URL from which the triple was extracted. Within Microdata, there are mainly two different vocabularies used: (1) *schema.org*, and (2) its predecessor *data-vocabulary.org*, which is deprecated since June 2011 [15].

[6]`http://www.w3.org/TR/microdata/`
[7]`http://any23.apache.org/`
[8]`http://www.w3.org/TR/n-quads/`

Table 1: Dataset statistics of filtered WebDataCommons Microdata datasets

WDC	# URLs	# PLDs	# Quads
2012	19 281 189	29 413	232 687 529
2013	217 751 199	399 139	6 411 276 458
2014	232 279 437	731 573	6 778 845 785

For this research, we are mainly interested in the usage and development of the schema.org vocabulary. Thus, we extracted only quads from the three original WebDataCommons Microdata datasets from 2012, 2013 and 2014 where the `predicate` or `object` includes the substring "schema.org". Although the vocabulary schema.org can potentially be used as well together with the markup language RDFa, previous studies have shown that only around 0.1% of all pay-level domains (PLDs) make use of this vocabulary together with RDFa [15]. Furthermore, JSON-LD, which is also possible to be used with schema.org, is not covered by the Common Crawl, and thus omitted in the corpus. Therefore, in the course of this research, we focus solely on schema.org data marked up using Microdata on HTML web pages.

Table 1 provides an overview of the size of the final datasets we used within our analysis. While the size of the crawls increases over the years, it is notable that from 2013 to 2014, the number of pages including schema.org Microdata is almost constant, while the number of PLDs increased by almost factor 2. This might be due to different PLDs contained in the two crawls.

As mentioned above, schema.org is driven by the four world-wide largest search engine companies, Google, Yahoo, Yandex and Microsoft (Bing). They maintain an active user/developer group which discusses and maintains the schema.org schema definition.[9] This community frequently creates whole new releases of the schema, where new classes and properties are introduced or domains and ranges of properties are changed. Also classes and properties are superseded by others or completely removed from the schema.[10]

We have extracted the RDF schema of the releases before and after the three crawls, i.e., release `0.91`,`0.93`, `1.0c`, `1.0f`, `1.91`, `1.92` and `1.93`, using the Internet Archive[11] for older releases, and the schema.org GitHub repository[12] for the newer ones. Figure 1 depicts the temporal order of the three used crawls and the analyzed schema.org changes between the selected release versions.

Table 2 shows the number of newly introduced classes and properties for each of the selected releases in comparison to the previous one, as well as the number of domain/range changes, deprecations, and supersessions.

In this table, S_i denotes the changes of the standard between the time when crawl i was started, and the time when crawl i was finished. These change sets are used to analyze top-down processes, i.e., adoptions of changes in the standard.

In contrast, S_i' denotes the changes of the standard between the *end* of crawl i and the beginning of crawl $i+1$.

[9]The most recent version of this definition can be found at the schema.org website `http://schema.org`.
[10]An overview about the major releases can be found at `http://schema.org/docs/releases.html`.
[11]`http://web.archive.org/`
[12]`https://github.com/schemaorg`

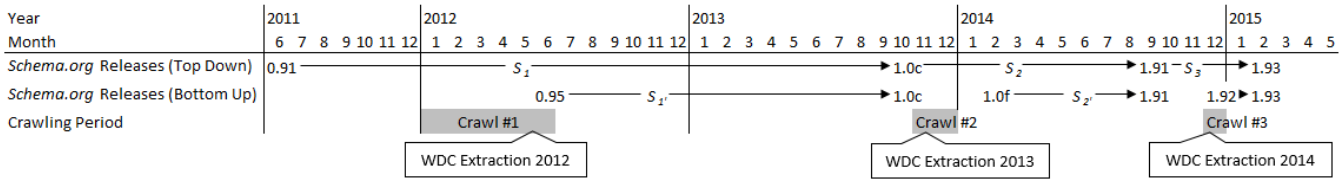


Figure 1: Timeline of schema.org release dates and crawl dates

These change sets are used to analyze bottom-up processes, i.e., influences of the deployed data on the standard.

3. RESEARCH QUESTIONS AND MEASURES

Generally, we base all of our measures on data aggregated by pay-level domains. This helps limiting the bias introduced by different crawling strategies used for the Common Crawl datasets [14, 15]. For defining the measures, we use the following notation conventions: our corpora are denoted with C_{2012}, C_{2013}, and C_{2014}, as explicated above. For each corpus C_i, t_i denotes the time at which it was collected, and $\#PLD_i$ denotes the total number of PLDs in the corpus deploying schema.org Microdata. Furthermore, for a triple pattern T, we define $\#PLD_i(T)$ as the total number of PLDs in a corpus which use the triple pattern T at least once.[13]

To quantify the usage of a class c and a property p, we define

$$\#PLD_i(c) := \#PLD_i(\texttt{?x rdf:type c}) \quad (1)$$

$$\#PLD_i(p) := \#PLD_i(\texttt{?x p ?y}) \quad (2)$$

as the total number of usages of types and properties aggregated by PLD.

3.1 Top-down Processes

Top-down processes are "schema first" processes, meaning that the standard changes and the data providers follow the standard. Here, we analyze how changes in the standard are reflected in the data *after* the change has been defined. More specifically, we look into:

- Adoption of new classes and properties
- Implementation of deprecations
- Implementation of domain/range changes

[13]We use the notion of triple patterns as defined in the W3C SPARQL standard [21].

Table 2: Overview of the different sets of changes between the selected Releases. We separate changes introducing new concepts and properties (new) and removing existing concepts and properties (dep).

Δ	Release from	Release to	# Classes new	# Classes dep	# Propert. new	# Propert. dep	# Domain /Range new	# Domain /Range dep	# super-session
S_1	0.91	1.0c	233	0	387	2	23	2	0
$S_{1'}$	0.95	1.0c	140	0	161	0	34	1	0
S_2	1.0c	1.91	62	0	190	1	119	9	32
$S_{2'}$	1.0f	1.91	36	1	108	1	32	5	32
S_3	1.91	1.93	27	0	76	1	68	69	3
$S_{3'}$	1.92	1.93	2	0	15	1	22	5	0

For measuring the adoption of a new schema element s (i.e., a class or a property), we determine the *normalized usage increase (nui)* of that element as

$$nui_{ij}(s) := \frac{\#PLD_i(s)}{\#PLD_j(s)+1} \Big/ \frac{\#PLD_i}{\#PLD_j} \quad (i > j) \quad (3)$$

The nominator of the overall fraction denotes the increase in the usage of s between the corpora C_i and C_j, whereas the the denominator denotes the general increase of deployed schema.org Microdata contained in the crawl. In order to avoid division by zero for elements that have not been used previously, use $\#PLD_j(s)+1$ as a denominator instead of $\#PLD_j(s)$.

The usage of a normalized measure is steered by the raw data which we use for our analysis. The underlying web crawls – based on nature of web crawls – do not include the same sets of web pages and do also not include the same number of crawled pages. By this, even if the total number of adopting sites could be larger, the relative amount could be smaller as in the crawl before. These facts forbid the usage of non-normalized scores such as the simple differences between the total number of pages adopting a particular class.

For a new schema element s added to the standard between two released t_i and t_{i+1}, we say that it has been successfully adopted if there is a $i > j$ so that $nui_{ij}(s) \geq 1.05$, i.e., the increase of the usage of the element is significantly larger than the overall increase in schema.org Microdata. Likewise, for deprecated elements, we say that the deprecation has been successfully adopted if $nui_{ij}(s) \leq 0.95$, i.e., the usage of the element has significantly decreased.

The rationale for the normalization is that, assuming there are no other influencing factors, it can be expected that the usage increase of an element increases proportionally with the overall increase of the corpus. Only if the usage increase of an element significantly *exceeds* this expected increase, we can say that there is a measurable impact of the change in the standard.

For domain and range changes, we have to distinguish between classes being *added* to the domain/range definition, and classes being removed. Due to the disjunctive interpretation of domain and range definitions in schema.org [17], the further *broadens* the possible usages of a property, while the latter *restricts* the possible usages, i.e., the latter can lead to formerly legal definitions to become illegal.

For measuring the adoption of domain changes of a property p and a domain d, we count triple patterns of the type

$$\texttt{?x p ?y . ?x rdf:type d'}, \quad (4)$$

where `d'` is a subtype of `d`, or `d` itself. For range changes with a range `r`, we count triple patterns of the type

$$\texttt{?x p ?y . ?y rdf:type r'}, \quad (5)$$

where `r'` is a subtype of `r`, or `r` itself. For such a patterns, we define $nui_{ij}(p)$ as in (3).

As for new classes and properties, we say that an addition to a domain/range definition is adopted if the corresponding $nui_{ij}(p) \geq 1.05$. We say that a removal from a domain/range definition is adopted if the corresponding $nui_{ij}(p) \leq 0.95$.

3.2 Bottom-up Processes

Bottom-up processes are "data first" processes, meaning that the standard is adapted to its actual implementations and deployments. Here, we analyze how changes in the standard are reflected in the data *before* the change has been defined. More specifically, we look into:

- The usage of (undefined) classes and properties before they were officially announced
- The adoption of schema.org's extension mechanism[14] to define new classes
- The usage of properties with subjects and objects not defined in their range

To measure whether there are such bottom-up processes, we hypothesize that elements that are already used by a larger number of data providers *prior to announcement* are likelier to be included in the standard. As a measure for testing this hypothesis, we use receiver operator characteristics, i.e., ROC curves [8].

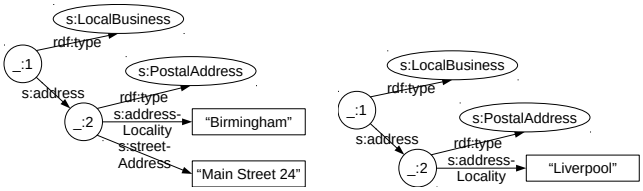
ROC curves are often used in measuring the performance of predictors, such as machine learning trained classifiers. Here, we measure if the number of PLDs which deploy a specific schema element is a good predictor for that element to become officially added to the standard.

The ROC curves are contructed as follows. Given two corpora C_i and C_{i+1}, let A_{i+1} denote the set of schema elements that have been *added* to the standard between t_i and t_{i+1}. Furthermore, let S_i be the list of undefined schema elements (according to the standard at time t_i) used in C_i, ordered by $\#PLD_i(s)$. Then, we mark each element in S_i as a *true positive* if it is also contained in A_{i+1}, as a *false positive* otherwise. Given the ordered list, we graph the true positive rate against the false positive rate, and measure the area under the curve (AUC), which is normalized to a $[0, 1]$ range. We build individual ROC curves for classes and properties.

If $AUC = 0.5$, then there is no influence of the usage of an element on the probability of it being included into the standard. For $AUC > 0.5$, there is a positive influence (i.e., more frequently deployed elements are likelier to be included into the standard later), if $AUC < 0.5$, there is even a negative influence.

Likewise, we analyze whether the usage of the extension mechanism has an influence on the standardization. For example, the class `s:Artwork` has been newly introduced in the recent schema.org release, being a subclass of `s:CreativeWork`. Before that official introduction, it could have been used via the extension mechanism by defining the class `s:CreativeWork/Artwork`, which then recognized as a user-defined subclass of `s:CreativeWork`. Like for inofficially used elements, we compare the list of schema elements defined in C_i using the extension mechanism to the corresponding set A_{i+1} of new schema elements in the standard at t_{i+1}, and compute ROC curves both for classes and properties.

[14] `http://schema.org/docs/extension.html`



(a) Document 1 (b) Document 2

Figure 2: Two example documents. Document 2 defines only a city, but no street.

To measure whether domain and range changes are influenced by the actual usage of data, we look specifically at domain and range definitions that have become *broader*. To that end, we look at all domain and range usages in a corpus C_i according to (4) and (5) which are not defined in the standard at t_i. Again, we sort them by $\#PLD_i(p)$ and mark all domain/range definitions that have been added to the standard at t_{i+1} as true positives, the rest as false positives. The resulting ROC curves show if there is a tendency to add domain and range definitions based on the deployed usage.

3.3 Overall Convergence of Vocabulary Usage

The third question we raise is the overall convergence or divergence of schema.org Microdata. Specifically, we want to know if the diversity of representing particular entities – such as an address – has increased or decreased over time. Convergence is a plausible scenario due to the increased availability of tutorials and best practices, e.g. for *Google Rich Snippets*[15], or adaption to the consumers of Microdata, such as optimization w.r.t. search engine rankings. On the other hand, divergence is also possible due to the larger number of adopters, all of which come from different domains and backgrounds with specific requirements.

To quantify convergence and diversity, we adapt a normalized entropy measure [20]. Since the RDF data extracted from a web page forms a cycle-free RDF graph with a defined set of roots [11], we first describe the *vocabulary usage* of the page as the ordered enumeration of all paths from any root to any leaf. For the paths, we extract the types and properties, but omit blank node identifiers and literal values. To enforce an ordering, we use a simple lexicographic ordering. For example, the two documents shown in Figure 2 would be described by the following enumerations:

$$S_1 = \{ \texttt{s:LocalBusiness} \rightarrow \texttt{s:address/s:PostalAddress}, \texttt{s:LocalBusiness} \rightarrow \texttt{s:address} \rightarrow \texttt{s:addressLocality}, \texttt{s:LocalBusiness} \rightarrow \texttt{s:address} \rightarrow \texttt{s:streetAddress} \} \quad (6)$$

and

$$S_2 = \{ \texttt{s:LocalBusiness} \rightarrow \texttt{s:address/s:PostalAddress}, \texttt{s:LocalBusiness} \rightarrow \texttt{s:address} \rightarrow \texttt{s:addressLocality} \} \quad (7)$$

The set of those enumerations is now treated as a sequence of symbols, where each enumeration element (i.e., each path)

[15] `https://developers.google.com/structured-data/rich-snippets/`

is understood as a symbol. Thus, we can compute the total entropy for the set of the two example documents as

$$H = \sum_{i=1}^{n} -p(path_i)log(p(path_i)), \quad (8)$$

In the above example, the total entropy would be 1.918, using the dual logarithm. We normalize this by dividing by the product of the total number of paths N in a corpus, and

$$H_{max} = log(n), \quad (9)$$

where n is the total number of *different* paths, to account for effects of different corpus sizes (i.e., we use a *normalized entropy rate*):

$$H_{norm} = \frac{H}{H_{max} \cdot N}. \quad (10)$$

In the above example, this would lead to a normalized entropy rate of 0.192.

If we now assume that at a point in time, the second document would also add a `s:streetAddress`, i.e., the two documents would become more alike, the normalized entropy rate would drop to 0.167. Thus, we can observe that the description of entities of the class `s:LocalBusiness` has become more uniform.

We compute an overall normalized entropy, as well as normalized entropies per class defined in schema.org.

3.4 Influence of Data Consumers

As stated above, one major incentive to directly markup concepts within the HTML pages is, among other things an improved display on the search result page, and by this a higher likeliness to attract users. Google, the most widely used search engine[16], calls those improved displays *Rich Snippets* and supports web site provides within their Google Developer Tools pages for structured data with How-Tos and examples, as discussed above. In particular Google promotes the seven different domains/classes *Products*, *Recipes*, *Reviews*, *Events*, *SoftwareApps*, *Videos*, and *Articles*, and explicitly states which properties are required for being properly displayed within the Rich Snippets. In our analysis, we will look at the measures for those classes in isolation, where appropriate.

In addition, one could assume that the introduction of copy-pasteable code snippets (e.g. examples) how to implement a certain description/markup language within HTML can boost the deployment. However, such examples have been available on the schema.org web site even before the first crawl we use, so we do not expect any significant findings from the availability such examples.

4. EMPIRICAL FINDINGS

In this section, we describe and analyze the empirical findings for top-down and bottom-up processes, as well as the overall convergence of the data descriptions.

4.1 Top-down Processes

Based on the timeline in Figure 1, for the following analysis we consider the set of changes S_1, including the changes between the releases `0.91` and `1.0c`, and the set of changes S_2 including the changes between the releases `1.0c` to `1.91`.

[16] `http://gs.statcounter.com/#search_engine-ww-monthly-200807-201503`

Table 3: Median and average of class and property *nui*-values.

Change Set	Classes		Properties	
	S_1	S_2	S_1	S_2
Median	0.00	0.55	0.12	0.00
Average	0.47	8.04	2.01	6.63

We make use of the three crawls to calculate the normalized usage increase (*nui*) between two datasets i and j for a schema element s based on the number of PLDs making use of this particular element.

4.1.1 Adoption of New Classes and Properties

Overall, we observe that from the 233 new classes introduced with S_1, 113 could *not be observed at all* in any of our corpora until 2014. Likewise, regarding the 389 new properties in S_1, 109 could not be observed until 2014. Within the change set S_2, 23 out of 62 newly introduced classes and 109 out of 190 newly introduced properties were not used at all in the 2014 corpus.

These findings show that the adoption of new classes and properties in general is happening slowly, and that there are certain parts of the schema which are barely used at all. This is also reflected in the average and median *nui*-values reported in Table 3. In particular, the low median values show that the vast majority of newly introduced classes and properties is not significantly adopted.[17]

By manually inspecting the lists of non-adopted classes and properties, we could identify three particular domains. For elements introduced in S_1, we could not find any evidence for parts of the objects (1) from the medical domain (e.g., `s:Nerve` or `s:Vein`), as well as (2) for many specific subclasses of `s:Action`. The first is an effect of integrating an existing large-scale, multi-purpose vocabulary for a domain – in this case, the medical domain – into the schema[18], where not all parts of that vocabulary are equally useful for marking up web page content.

The cause for the latter may be a blind spot of our corpora, because actions are basically designed for e-mail markup, not web-page markup[19]. For elements introduced in S_2, the main domain of non-adopted elements are related to booking actions – here, again, the markup is likely to be used in confirmation e-mails and form-based interaction with the deep web, both of which are not included in the Common Crawl.

Table 4 lists the 19 significant deployed classes of the change sets S_1 and S_2, which are at least deployed by five PLDs in the 2014 dataset based their *nui*-value. Although we have found a large fraction of action-related and medical-related classes beyond those not being adopted at all (see above), two classes from those domains, i.e., `s:SearchAction` and `s:MedicalIndication`, are listed in the table. Within S_2, we find a large fraction of classes related to the broadcasting domain, as well as services. In addition the FAQ-related classes are present, like `s:Question` and `s:Answer`, as

[17] A median of 1.05 would confirm that half of the classes/properties are significantly adopted.

[18] `http://blog.schema.org/2012/06/health-and-medical-vocabulary-for.html`

[19] `https://developers.google.com/gmail/markup/`

Table 4: 19 significant deployed classes of S_1 and S_2 between 2013 to 2014 being at least used by 5 PLDs in 2014 and the 7 classes directly promoted by Google Developer Tool web page for Rich Snippet integration.

Δ	Class	#PLD '13	#PLD '14	*nui*
S_1	s:SearchAction	1	11	3.00
S_1	s:MedicalIndication	1	5	1.36
S_1	s:BusinessFunction	2	7	1.27
S_2	s:WebSite	2	1 648	299.71
S_2	s:Car	0	76	41.46
S_2	s:QAPage	0	60	32.74
S_2	s:Answer	0	41	22.37
S_2	s:Question	1	47	12.82
S_2	s:PublicationIssue	0	21	11.46
S_2	s:Vehicle	0	21	11.46
S_2	s:Periodical	0	16	8.73
S_2	s:BroadcastEvent	0	14	7.64
S_2	s:BroadcastService	0	14	7.64
S_2	s:Episode	1	17	4.64
S_2	s:Service	18	147	4.22
S_2	s:EmailMessage	4	35	3.82
S_2	s:ServiceChannel	0	5	2.73
S_2	s:Airline	0	5	2.73
S_2	s:RadioEpisode	2	7	1.27
	s:Product	56 537	89 683	0.87
	s:Recipe	6 025	7 593	0.69
	s:Review	13 143	20 115	0.84
	s:Event	8 253	10 105	0.67
	s:SoftwareApplication	1 809	2 091	0.63
	s:VideoObject	4 516	7 424	0.90
	s:Article	65 864	88 569	0.73

schema.org has been adopted by major question-and-answer sites such as *Stack Overflow*[20].

Furthermore, Table 4 lists the seven classes which are promoted by Google's Developer Tools in order to mark content for *Rich Snippets*. Although the *nui*-value is below 0.95, the absolute number of PLDs using those classes is still growing, but not as fast as the overall deployment of schema.org for Microdata. Especially the classes `s:VideoObject` (64%), `s:Product` (58%), and `s:Review` (53%) have strongly increased in total number of deployment by PLDs.

Regarding the properties introduced in S_1 and S_2, we found 13 respectively 56 being significantly deployed within then 2014 corpus. Table 5 lists the 40 significant deployed properties of both change sets based on their *nui*-value, which are deployed by at least 5 PLDs. We added the possible domains for all of the properties in order to allow a straightforward grouping by topical domain.

A first observation, which we could already draw from the significant deployed classes (see above), is that a large fraction of those significantly deployed properties are only used by a small number of PLDs, especially for S_1. Regarding S_2, we see a stronger deployment by number of PLDs. This is remarkable, since time is apparently not a crucial factor in adoption, i.e., elements that have been present in the standard for a longer time are not necessarily adopted more widely. Similar to the classes, most of the properties have domains of the groups: *Action*, *MedicalEntity*, *CreativeWork*, *ContactPoint*, *Organization*, *Service*, *Question* and *Events*.

In order to gather additional insights and try to identify groups of classes which define significantly properties we calculated, based on the properties within S_1 and S_2 for the comparison of 2013 and 2014 the average *nui*-values for all possible classes. For this calculation we exclude all properties which are inherited from the class `s:Thing`, to reduce the noise from too general properties.

[20] `http://stackoverflow.com/`

Table 5: 42 significant deployed properties of S_1 and S_2 between 2013 to 2014, being deployed at least by 5 PLDs in 2014.

Δ	Property	Domains (Excerpt)	#PLD 13	#PLD 14	*nui*
S_1	s:result	Action	2	10	1.82
S_1	s:agent	Organization	1	6	1.64
S_1	s:endTime	Action	2	7	1.27
S_1	s:object	Action	3	9	1.23
S_1	s:codeValue	MedicalCode	4	11	1.20
S_1	s:medicineSystem	MedicalEntity	3	8	1.09
S_2	s:potentialAction	Thing	0	783	427.20
S_2	s:target	Action	0	783	427.20
S_2	s:commentCount	CreativeWork	0	98	53.47
S_2	s:hasMap	Place	0	54	29.46
S_2	s:contactOption	ContactPoint	1	101	27.55
S_2	s:doorTime	Event	1	80	21.82
S_2	s:pagination	Article	0	32	17.46
S_2	s:department	Organization	7	256	17.46
S_2	s:acceptedAnswer	Question	0	29	15.82
S_2	s:position	CreativeWork, ListItem	0	27	14.73
S_2	s:subOrganization	Organization	4	121	13.20
S_2	s:suggestedAnswer	Question	0	23	12.55
S_2	s:partOfSeries	Episode, Season	0	22	12.00
S_2	s:organizer	Event	2	65	11.82
S_2	s:areaServed	ContactPoint	0	21	11.46
S_2	s:answerCount	Question	0	18	9.82
S_2	s:productSupported	ContactPoint	1	34	9.28
S_2	s:upvoteCount	Anser, Comment, Question	0	17	9.28
S_2	s:serviceArea	Service	2	48	8.73
S_2	s:serviceType	Service	4	73	7.97
S_2	s:audienceType	Audience	0	13	7.09
S_2	s:accessibility-Feature	CreativeWork	0	12	6.55
S_2	s:issueNumber	PublicationIssue	0	12	6.55
S_2	s:availableLanguage	ContactPoint, ServiceChannel	0	11	6.00
S_2	s:mapType	Map	0	11	6.00
S_2	s:publishedOn	PublicationEvent	0	10	5.46
S_2	s:produces	Service	1	19	5.18
S_2	s:numberOfSeasons	*Series	0	9	4.91
S_2	s:directors	Episode, Movie	0	7	3.82
S_2	s:hasPart	CreativeWork	0	6	3.27
S_2	s:eventStatus	Event	2	17	3.09
S_2	s:hoursAvailable	ContactPoint	3	22	3.00
S_2	s:reservationFor	Reservation	0	5	2.73
S_2	s:publication	Clip, Episode, MediaObject	4	12	1.31
S_2	s:issn	Periodical	3	8	1.09
S_2	s:license	CreativeWork	17	36	1.09

The average values for the selected classes are displayed in Table 6. The table ranks all the classes we have identified earlier and all of them – except for `s:MedicalEntity` – have an average *nui*-value above the significance level.

4.1.2 Implementation of Deprecations

With the changes of S_1, no deprecations were introduced. Within S_2 one property became completely deprecated, but it was not used in any of the three crawls. Beside the complete deprecation of one property, 32 became superseded by others. Out of those 32, the usage of superseded properties significantly decreased for 29 of those, except for the three properties `s:map`,`s:maps` and `s:musicGroupMember`, where `s:map` is still significantly used in 2014. Major users of the `s:map` properties are for example hotel site like `marriott.com`, travel sites and blogs like `travelpod.com`.

Table 6: Excerpt of classes ordered by the calculated average *nui* based on properties from S_1 and S_2 for the 2013 and 2014 crawl.

Rank	Class	Avg. *nui*
1	s:Action	48.61
... other `s:Action`-subclasses		
64	s:PostalAddress	11.46
65	s:ContactPoint	11.46
66	s:Service	5.14
66	s:Taxi	5.14
68	s:Question	4.94
69	s:Event	4.28
... other `s:Event`-subclasses		
92	s:TVSeason	4.03
... other TV-related-subclasses		
... other mixed classes		
126	s:Places	3.41
... other `s:Place`-subclasses		
... other mixed classes		
181	s:Organization	1.62
... other `s:Organization`-subclasses		
280	s:MedicalEntity	0.82
... other `s:MedicalEntity`-subclasses		
... other classes		

Table 7: Significantly used substitutes of superseded properties of S_2 within the 2014 dataset. In all those cases, the property supersedes the respective property named by the plural form, i.e., `s:blogPost` supersedes `s:blogPosts`, etc.

Property	# PLDs'13	# PLDs'14	*nui*
s:blogPost	5 445	33 946	3.40
s:employee	214	745	1.90
s:member	254	871	1.87
s:sibling	3	9	1.64
s:event	149	369	1.35
s:award	102	235	1.26
s:contactPoint	331	726	1.20
s:season	42	85	1.10
s:photo	1 004	1 962	1.07

From the substitutes for the superseded properties which should be used after the changes of S_2, we could find nine to be adopted significantly within the 2014 crawl, listed in Table 7. For the remaining 23, there is no significant adoption for the substitutes.

4.1.3 Implementation of Domain/Range Changes

Based on our observations, none of the range changes of properties which were introduced within S_1 is significantly deployed in any of the three corpora. From the 18 introduced domain changes in S_1, six are adapted by a significant amount of PLDs in the later corpora. The adoptions for those changes which are deployed by at least five PLDs in 2014 are listed in Table 8. Four are directly related to the *product* domain.

From the 12 significant deployed domain changes (out of 87) Table 8 lists the eight which are used by at least five PLDs in 2014. In addition the table also includes the seven adopted range changes (out of 20) included in S_2. A large proportion of those adoptions can be assigned to the *broadcasting* domain. That domain was introduced into the schema.org vocabulary based on discussions and influence with BBC and EBU.[21]

4.2 Bottom-up Processes

[21] `http://blog.schema.org/2013/12/schemaorg-for-tv-and-radio-markup.html`

Table 8: List of domain/range changes significantly adopted and at least deployed by 5 PLDs in 2014. (+) indicates a new range/domain, (−) the removal of an range or domain.

Δ	Change	# PLDs '13	# PLDs '14	*nui*
Domain				
S_1	s:Product/width (+)	99	318	1.73
S_1	s:Product/itemCondition (+)	360	1187	1.79
S_1	s:Drug/manufacturer (+)	13	32	1.25
S_1	s:PriceSpecification/priceCurrency (+)	100	215	1.16
S_1	s:Product/height (+)	85	299	1.90
S_2	s:Event/typicalAgeRange (+)	0	6	3.27
S_2	s:Organization/memberOf (+)	1	5	1.36
S_2	s:TVEpisode/episodeNumber (−)	75	106	0.76
S_2	s:Thing/alternateName (+)	1	14	3.82
S_2	s:Service/provider (+)	2	55	10.00
S_2	s:WebPage/isPartOf (−)	43	68	0.84
S_2	s:RadioSeries/episode (+)	0	5	2.73
S_2	s:Episode/actor (+)	1	7	1.91
Range				
S_2	s:comment|s:Comment (+)	44	172	2.13
S_2	s:seasons|s:TVSeason (−)	9	8	0.48
S_2	s:episodes|s:TVEpisode (−)	11	14	0.69
S_2	s:partOfSeason|s:TVSeason (−)	15	22	0.80
S_2	s:isPartOf|s:CollectionPage (−)	18	30	0.91
S_2	s:episode|s:TVEpisode (−)	56	78	0.76
S_2	s:image|s:ImageObject (+)	101	264	1.43

In this section, we report on the numbers of classes, properties and other changes which are actually adopted by web pages before they became official within the schema definition of schema.org. In particular we inspect the changes made starting from release `0.95` for the first crawl ($S_{1'}$, S_2, and S_3), and from release `1.0f` for the second crawl ($S_{2'}$ and S_3) and from release `1.93` to the current release for the last crawl ($S_{3'}$). We are aware of the fact that before a change is officially announced, there are ongoing discussions and proposals (which are all public), which also could affect the earlier deployment of non-official classes, and will take this into account when drawing any conclusions.

4.2.1 Usage of Classes and Properties before Official Announcement

Regarding the usage of (undefined) classes and properties within the deployed data before they were officially included in the standard, we can report in general a rather small pre-announcement deployment. From the changes of S_1' we identified only one class and 13 properties being already deployed in the crawl of 2012. The most deployed properties were `s:value` and `s:color` which were both used by four different PLDs.

Analyzing the influence of the deployment in 2012 and 2013 for the changes until release `1.91`, we found that within the first crawl only the class `s:Service` was deployed by one PLD and three other properties were already present. Within the data of 2013 we found four classes and eight properties being deployed. Those mainly belong to the domain of flights, where the class `s:Flight` was deployed by six different PLDs together with their properties `s:iataCode`, `s:arrivalAirport`, and `s:departureAirport`. Those are not big airlines, but copies of one and the same meta-flight booking portal: `aviagid.com.ua`.

Regarding the influence of the deployed classes and properties for the `1.93` release, we found that the class `s:Game` was already used in 2012 by six PLDs, and by 18 PLDs in 2013 before it got officially released. We also found three

Table 9: AUC values for bottom up adoption of classes, properties, and domain and range changes between the different datasets.

	2012/2013	2012/2014	2013/2014	AVG.
Classes	0.4272	0.3739	0.7305	0.5105
Properties	0.5369	0.5292	0.8547	0.6403
Domain Changes	0.7449	0.7449	–	0.7449
Range Changes	–	0.9498	0.9827	0.9662

respectively six properties being deployed in 2012 and 2013, with the property `s:currency` being used by 24 PLDs in 2012 and already 551 in 2013 is most outstanding. Within the crawl of 2014, we could identify the property `s:material` being already used by six PLDs before the official release.

As described in Section 3.2, we draw the ROC curves for the three dataset comparisons and calculate the corresponding AUC values. Figure 3 shows the different curves for the three comparisons for classes (black line) and properties (black dotted line). As stated above, for the classes, we could only identify one, respectively two classes which are used before the official release, which explains the angular curves. For the properties we could find more adoptions, but the curve also follows more or less the diagonal.

Table 9 shows the calculated AUC values for the comparisons based on the ROC curves of Figure 3. The values for classes and properties for the comparison with 2012 show a more or less random distribution, where the comparison of 2013 and 2014 shows a stronger trend towards an influence of pre-official usage of classes and properties. Summarizing the influence, based on the average in the last column of this table, shows a minor trend overall.

4.2.2 Adoption of schema.org's Extension Mechanism

When looking for new classes and properties being used via the extension mechanism before their official introduction, we found only three class extensions in the 2012 corpus (`s:*/Service`, `s:*/Vehicle`, and `s:*/WebApplication`), being used by maximum of two PLDs. Furthermore, ten properties are introduced using this mechanism, with `s:*/softwareVersion` being used most (by five PLDs).

For the 2013 corpus, three properties were used with the extension mechanism and become later official. But the usage is always less than four PLDs. Class-wise, we again find `s:*/Vehicle` being deployed using the extension mechanism by nine PLDs, and seven further classes.

In 2014 we can report one class and five properties being introduced using the extension mechanism. Outstanding, again, is `s:*/currency`, which was used by ten PLDs.

Overall, regarding the extension mechanism, we cannot report any significant influence on the newly introduced properties and classes. In general, the mechanism is not widely used, and we can observe that data providers are more likely to introduce classes and properties directly without using the official extension mechanism.

4.2.3 Usage of Properties with Subjects and Objects Outside their Defined Domain and Range

In addition to classes and properties, we analyzed the pre-official usage of domains and ranges with properties, where the domain/range was not defined yet at the point in time of the crawl. In other words, we look at properties being used in a different *context* than the one they were intended to be used.

Overall, we found that six domain/range changes (four domain and two range changes) can be detected within the crawled data before they become official. Especially the range changes `s:comment` with its new range `s:Comment` and `s:image` with its new range `s:ImageObject` are already used by over 40 and 100 PLDs, respectively, in the 2013 data. A prominent example for using a property with a new domain is `s:Organziation/brand`, which was already present on 255 PLDs in 2012 although it was not official released.

We again draw the ROC curves as described in Section 3 and display the different curves for domain and range changes within Figure 3. From those curves and the corresponding AUC values, depicted in Table 9, we can observe that at least for domain and range changes, the schema evolution is driven by the real world usage to a certain extent, as the AUC values of those changes are significantly larger than 0.5.

4.3 Overall Convergence of Vocabulary Usage

To complete the picture of the evolution of deployed data over time, we had a look at the development of the heterogeneity of the usage of the different class definitions and also of the global data space.

As described in Section 3.3, we use an entropy-based measure for measuring heterogeneity. From an overall point of view, we find that the global normalized entropy rate, as defined in (10), and hence the heterogeneity, decreased from 2012 ($2.34e^{-09}$) to 2014 ($9.42e^{-11}$) by around 2 400%, i.e., we can observe a strong homogenization of the data representations.

Regarding the class-wise entropy and its development from 2012 to 2014, we have a closer look at the 57 most deployed classes (classes which we could find on more than 1 000 PLDs in the 2014 crawl)[22], the entropy decreases for 56. Only the entropy for the class `s:VideoObject` increased by around 18%. Comparing only the class-wise entropy for the 2013 and the 2014 crawl, we can report that 37 increase in homogeneity and 17 decrease.

Table 10 list those 37 classes for which we found a decrease of their class-specific entropy from 2013 to 2014. The classes listed here can be categorised in four different categories:

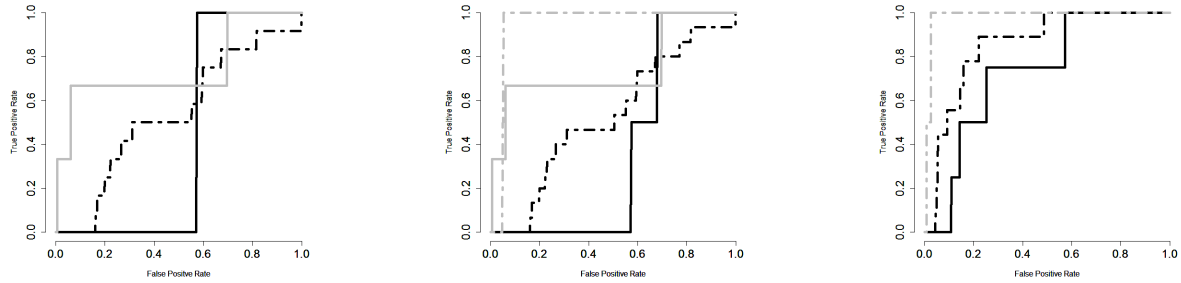
1. Classes describing web sites and their elements and structured like `s:WebSite`, `s:ImageGallery`, `s:Blog`, `s:WPSidebar`, `s:WPHeader`. The increase in homogeneity of such kind of classes can be explained by the increasing deployment of schema.org markup language within Content Management Systems.[23]

2. Services and facilities, like `s:Florist`, `s:AutoDealer`, `s:Hotel`, `s:Restaurant`, and `s:Store`, mostly belonging to the class of `s:LocalBusiness`. Those classes are mainly deployed by yellow-pages websites.

3. Products and offers, like `s:Product`, `s:ItemList`, and `s:Offer`. Here the promotion of Googles Developer Tools and Rich Snippets could be a possible driver.

[22]For this experiment we focus only on classes which were already deployed in 2012.

[23]For example the CMS Drupal (starting with Version 7) automatically annotates the generated pages within their system using schema.org classes and properties: `https://www.drupal.org/project/schemaorg`.



(a) 2012/2013 Bottom Up Dataset Comparison

(b) 2012/2014 Bottom Up Dataset Comparison

(c) 2013/2014 Bottom Up Dataset Comparison

Figure 3: ROC for each dataset comparison for classes (black line), properties (black dotted line), domain changes (grey line), and range changes (grey dotted line)

4. Ratings and reviews, which can be found in all of the three categories above: `s:Rating` and `s:Review`.

At the other end of the spectrum, Table 11 lists the 17 classes with a decrease in homogeneity between 2013 and 2014. Here, we can observe two high-level classes, i.e., `s:LocalBusiness` and `s:Event`, for which a larger number of specific sub classes have been introduced in later revisions of schema.org, so that instances of those classes have become richer (and more diverse) in their descriptions.

Another group of classes with increasing heterogeneity are classes describing locations, like, e.g., `s:PostalAddress`, `s:Place`, `s:GeoCoordinates`. As shown in [14], those classes were mostly used erroneously in the 2013 crawl, thus, the change to a more "correct" representation might lead to this decrease (i.e., the descriptions get more heterogeneous since correct and incorrect representations are used side by side).

5. RELATED WORK

There exists a large body work examining the adoption of different technical standards [3, 4, 5, 7, 22] as well as the role of standardization bodies in this process [6]. Example studies investigating the factors that drive the adoption of electronic data interchange (EDI) standards include [3, 4, 22]. Studies that focus on the adoption of Web service technologies include [5, 7].

The work presented in this paper distinguishes itself from and enriches the existing body of work on standard adoption by, on the one hand side, investigating the adoption of a different, rather recent pseudo-standard, i.e., the schema.org vocabulary. On the other hand, we apply a different research methodology: instead of using questionnaires that are sent to a likely incomplete set of potential adopters, with a rather small number of responses, we analyze a series of large-scale Web crawls in order to cover a large fraction of the target population of all existing websites.

More closely related work on the adoption of specific vocabularies for publishing structured data on the Web includes Ashraf et al. [1] and Kowalczuket al. [13] which both investigate the adoption of the GoodRelations vocabulary for representing e-commerce data. A study of the adoption of the Web Ontology Language (OWL) is presented by Glimm et al. [9]. This paper distinguishes itself from these work by on the one hand focusing on a different vocabulary and on the other hand by analyzing the diffusion of the vocabulary over a longer time span. The adoption of metadata and license vocabularies for Linked Open Data is analyzed, e.g., [12] and [19].

The adoption of the Microdata, RDFa, and Microformat syntaxes for annotating structured content in web pages is investigated by Mika and Potter [16] as well as in two of our previous papers [2, 15]. The study presented in this paper distinguishes itself from these previous works by not only reporting the rise in overall adoption but also investigating the factors that potentially drive this adoption.

Patel-Schneider [17] analyzes the design of the schema.org vocabulary from an ontology modeling perspective and proposes changes to the overall design of the vocabulary in order to align it more closely with existing knowledge representation standards, in particular the Web Ontology Language (OWL).

6. CONCLUSION AND OUTLOOK

In this paper, we have shown how the availability of data deployed on the web using a given standard, and the standard itself, allows for a new kind of empirical analysis of standard adoption, which is completely data-driven. Using snapshots of the Web at different points in time, we are able to observe processes in the adoption of a standard, as well as its evolution. Such observed processes, as well as the identification of key drivers and obstacles in standard adoption are also useful insights for the adoption of other standards, such as Linked Open Data [18].

For this paper, we have focused on schema.org and its deployment using Microdata. To that end, we have taken a diachronic perspective, comparing snapshots of deployed schema.org Microdata to the corresponding versions of the standards.

The findings from our quantitative analysis are manifold. First, by far not all elements introduced in the schema.org standard are actually deployed – in fact, about half of the defined elements could not be observed in any of the corpora. On the other hand, deprecations in the standard are most often adopted quite well.

On the other hand, we have also shown that bottom-up processes influence the evolution of the schema. In particular, the usage of defined properties in new contexts (i.e., with other domains and ranges than defined in the schema) often leads to corresponding changes in the schema. More-

Table 10: Classes with an increase of homogeneity from 2013 to 2014. Column 2 of this table states the class where the third column reports the change of the entropy. 100% in this column means, that the current (2014) entropy is only half of the entropy in 2013. Asterisks marks the promoted classes by Google Developer Tools.

Rank	Class	Increase of Homogenity in %	#PLDs in WDC 2014
1	s:WebSite	> 1 000.00	1 650
2	s:Thing	> 1 000.00	79 967
3	s:SiteNavigationElement	> 1 000.00	9 540
4	s:ImageGallery	> 1 000.00	1 679
5	s:RealEstateAgent	929.61	2 133
6	s:Florist	883.75	1 571
7	s:ItemList	582.59	1 697
8	s:Blog	422.45	110 531
9	s:IndividualProduct	378.50	1 403
10	s:WebPage	302.52	148 710
11	s:UserComments	251.87	9 128
12	s:AutoDealer	201.03	7 860
13	s:OpeningHours-Specification	155.96	1 163
14	s:Book	116.27	1 674
15	s:Dentist	114.03	2 410
16	s:SearchResultsPage	104.54	1 123
*17	s:Product	98.04	89 579
18	s:Movie	81.87	2 171
*19	s:Recipe	70.64	7 578
20	s:Corporation	65.44	1 900
21	s:CollectionPage	63.61	2 127
22	s:Offer	49.94	62 828
23	s:NutritionInformation	49.30	1 274
24	s:Brand	49.06	2 486
25	s:BlogPosting	41.36	65 320
26	s:ItemPage	30.82	3 455
27	s:WPSideBar	29.77	6 980
*28	s:VideoObject	28.06	7 419
29	s:Hotel	24.66	4 722
*30	s:SoftwareApplication	17.45	2 087
31	s:Rating	15.34	12 183
*32	s:Review	14.08	20 107
33	s:JobPosting	8.64	2 838
34	s:Store	5.94	1 819
35	s:Restaurant	4.26	2 524
*36	s:Article	1.16	88 164
37	s:WPHeader	0.56	7 879

over, we found that the intended way of introducing new classes and properties, i.e., the usage of schema.org's extension mechanism, is much less used than the (unintended) direct deployment of a new class or property.

In addition, we have also looked at the overall homogeneity of schema.org documents. We can observe that the homogeneity increases, e.g., (a) when there is a global player consuming the corresponding data, such as Google with its Rich Snippets for search results enrichment, or (b) by adoption of schema.org in widely deployed content management systems.

While the analysis in this paper is purely driven by the deployed data, we expect more fine-grained insights in standardization and adoption processes when extending the analysis to the mailing list archives where changes in the standard are often extensively discussed prior to changes to the standard[24], as well as the issue tracker on the schema.org GitHub, which is also used for feature requests, as well as for reporting, e.g., inconsistencies[25].

[24] `https://lists.w3.org/Archives/Public/public-vocabs/`

[25] `https://github.com/schemaorg/schemaorg/issues`

Table 11: Classes with an decrease of homogeneity from 2013 to 2014. Asterisks marks the promoted classes by Google Developer Tools.

Rank	Class	Decrease of Homogenity in %	#PLDs in WDC 2014
1	s:ProfessionalService	88.73	1 197
2	s:LocalBusiness	74.36	62 131
3	s:NewsArticle	73.98	2 514
4	s:ProfilePage	66.49	3 377
5	s:PostalAddress	65.14	100 960
*6	s:Event	57.57	10 091
7	s:MusicGroup	43.84	2 010
8	s:Place	30.87	9 912
9	s:AggregateOffer	30.69	2 038
10	s:Person	28.02	47 868
11	s:ApartmentComplex	26.60	1 921
12	s:ImageObject	20.67	25 529
13	s:CreativeWork	12.97	6 226
14	s:WPFooter	12.73	8 440
15	s:ContactPoint	8.42	1 034
16	s:Organization	4.61	52 658
17	s:GeoCoordinates	2.60	9 939

In summary, we have shown that the availability of both an open standard and web-scale data corpora can facilitate a novel type of rich empirical studies of standardization and adoption processes. The methods introduced in this paper can also be applied to different web standards.

7. ACKNOWLEDGEMENTS

The authors would like to thank Anna Primpeli for her assistance in extracting the latest Web Data Commons data corpora. The extraction has been supported by an Amazon Web Services in Education Grant award. Furthermore, we would like to thank the Common Crawl Foundation for providing the raw data on which the analyses in this paper are based.

8. REFERENCES

[1] J. Ashraf, R. Cyganiak, S. O'Riain, and M. Hadzic. Open ebusiness ontology usage: Investigating community implementation of goodrelations. In *LDOW*, 2011.

[2] C. Bizer, K. Eckert, R. Meusel, H. Mühleisen, M. Schuhmacher, and J. Völker. Deployment of rdfa, microdata, and microformats on the web – a quantitative analysis. In *ISWC*. Springer, 2013.

[3] P. Y. Chau and K. L. Hui. Determinants of small business edi adoption: an empirical investigation. *Journal of Organizational Computing and Electronic Commerce*, 11(4):229–252, 2001.

[4] M. Chen. Factors affecting the adoption and diffusion of xml and web services standards for e-business systems. *International Journal of Human-Computer Studies*, 58(3):259–279, 2003.

[5] M. Chen. An analysis of the driving forces for web services adoption. *Information Systems and e-Business Management*, 3(3):265–279, 2005.

[6] B. Chiao, J. Lerner, and J. Tirole. The rules of standard-setting organizations: an empirical analysis. *The RAND Journal of Economics*, 38(4):905–930, 2007.

[7] A. P. Ciganek, M. N. Haines, and W. Haseman. Horizontal and vertical factors influencing the

adoption of web services. In *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*, volume 6, pages 109c–109c. IEEE, 2006.

[8] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[9] B. Glimm, A. Hogan, M. Krötzsch, and A. Polleres. OWL: yet to arrive on the web of data? *CoRR*, abs/1202.0984, 2012.

[10] K. Goel, R. V. Guha, and O. Hansson. Introducing rich snippets. `http://googlewebmastercentral.blogspot.de/2009/05/introducing-rich-snippets.html`, 2009.

[11] I. Hickson, G. Kellogg, J. Tennison, and I. Herman. Microdata to rdf – second edition, 2014. `http://www.w3.org/TR/microdata-rdf/`.

[12] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the pedantic web. In *Linked Data on the Web*, 2010.

[13] E. Kowalczuk, J. Potoniec, and A. Ławrynowicz. Extracting usage patterns of ontologies on the web: a case study on goodrelations vocabulary in rdfa. In *OWLED*, 2014.

[14] R. Meusel and H. Paulheim. Heuristics for fixing errors in deployed schema.org microdata. In *Extended Semantic Web Conference*, 2015. to appear.

[15] R. Meusel, P. Petrovski, and C. Bizer. The webdatacommons microdata, rdfa and microformat dataset series. In *ISWC*, 2014.

[16] P. Mika and T. Potter. Metadata statistics for a large web corpus. In *LDOW 2012: Linked Data on the Web*, CEUR Workshop Proceedings, Vol. 937. CEUR-ws.org, 2012.

[17] P. F. Patel-Schneider. Analyzing Schema.org. In *International Semantic Web Conference*, 2014.

[18] H. Paulheim. What the adoption of schema.org tells about linked open data. In *2nd International Workshop on Dataset PROFIling & fEderated Search for Linked Data*, 2015.

[19] M. Schmachtenberg, C. Bizer, and H. Paulheim. Adoption of the linked data best practices in different topical domains. In *ISWC*, 2014.

[20] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[21] W3C. Sparql query language for rdf, 2008. `http://www.w3.org/TR/rdf-sparql-query/`.

[22] A. Yee-Loong Chong and K.-B. Ooi. Adoption of interorganizational system standards in supply chains: an empirical analysis of rosettanet standards. *Industrial Management & Data Systems*, 108(4):529–547, 2008.