What the Adoption of schema.org Tells About Linked Open Data

Heiko Paulheim

University of Mannheim, Germany
Research Group Data and Web Science
heiko@informatik.uni-mannheim.de

Abstract. schema.org is a common data markup schema, pushed by large search engine providers such as Google, Yahoo!, and Bing. To date, a few hundred thousand web site providers adopt schema.org annotations embedded in their web pages via Microdata. While Microdata and Linked Open Data are not 100% the same, there are some commonalities which make a joint analysis of the two valuable and reasonable. Profiling this data reveals interesting insights in the ways a schema is used (and also misused) on a large scale. Furthermore, adding a temporal dimension to the analysis can make the interaction between the adoption and the evolution of the standard visible. In this paper, we discuss our group's efforts to profile the corpus of deployed schema.org data, and suggest which lessons learned from that endeavour can be transferred to the Linked Open Data community.

Keywords: Microdata, schema.org, Linked Open Data, Data Profiling

1 Microdata and schema.org in a Nutshell

Microdata is a mechanism for embedding meta information in HTML.[1] Among its competitors, i.e., microformats[2] and RDFa[3], it is currently the most deployed annotation format [5].

Microdata is directly embedded into the HTML code. That is, different sections in the HTML code are marked up or annotated with schema classes and properties. In the following example, an address in an HTML page is marked up with Microdata:

```
<div itemscope itemtype="http://schema.org/PostalAddress">
 <span itemprop="name">Data and Web Science Group</span>
 <span itemprop="addressLocality">Mannheim</span>,
 <span itemprop="postalCode">68131</span>
 <span itemprop="addressCountry">Germany</span>
</div>
```

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

A parser like Any23[4] can extract the knowledge encoded in this HTML page, e.g., to RDF. The corresponding RDF triples in this example are:

```
_:1 a <http://schema.org/PostalAddress> .
_:1 <http://schema.org/name> "Data and Web Science Group" .
_:1 <http://schema.org/addressLocality> "Mannheim" .
_:1 <http://schema.org/postalCode> "68131" .
_:1 <http://schema.org/adressCounty> "Germany" .
```

Although it is possible to use arbitrary vocabularies for Microdata markup, schema.org[5] has become a de facto standard, with other vocabularies playing only minor roles, This is mainly due to the fact that schema.org is pushed my major search engines, i.e., Google, Yahoo!, Bing, and Yandex. While schema.org can be used both with Microdata and RDFa, the latter is only rarely deployed [1, 5].[6] In its latest release (1.93), schema.org comprises 620 classes and 890 properties.

2 Microdata and schema.org vs. Linked Open Data

As shown above, Microdata markup can be parsed into RDF, and thus, like RDFa, provides a means to publish Linked Data [2]. However, there are a few essential differences between Microdata and Linked Data, as it is commonly used.

Microdata, as shown above, is embedded into HTML. Since HTML documents themselves are trees, each the graph encoded by the RDF document extracted from Microdata is a *set of trees*. This means that Microdata is less expressive than pure RDF, which also allows for any *directed graphs* (containing cycles, and even more advanced constructs such as reification).

In 2006, Tim Berners-Lee formulated four principles for publishing Linked Data, i.e.[7]

1. Use URIs as names for things,
2. Use HTTP URIs so that people can look up those names,
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL), and
4. Include links to other URIs. so that they can discover more things.

In the example above, blank nodes were used for identifying concepts annotated on the web site, following the W3C recommendation[8]. Thus, in that case, URIs are not suitable names for things, as blank node identifiers are volatile, and neither are they resolvable by HTTP requests.

[4] `https://any23.apache.org/`
[5] `http://schema.org`
[6] Furthermore, although possible, schema.org is rarely used in Linked Open Data.
[7] `http://www.w3.org/DesignIssues/LinkedData.html`
[8] `http://www.w3.org/TR/microdata/`

As far as links to other resources are concerned, schema.org foresees the property `sameAs`[9], which, however, is currently deployed by less than 0.02% of all Microdata providers.[10] Thus, in its current form, schema.org Microdata only fulfills the third out of the four principles, if we accept Microdata as a standard on equal terms with RDFa.

In the same document, Berners-Lee created the *five star scheme* in 2010, defining five levels of Linked Open Data:

* Available on the web (whatever format) but with an open licence, to be Open Data
** Available as machine-readable structured data (e.g. excel instead of image scan of a table)
*** as (2) plus non-proprietary format (e.g. CSV instead of excel)
**** All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
***** All the above, plus: Link your data to other peoples data to provide context

While the license issue is tricky (many web pages do not come with an explicit license for their content), the first four stars are fulfilled by Microdata.

These reflections show that while Microdata is not essentially the same as Linked Open Data, there are a few commonalities which render it reasonable to have a closer look at both together, and see what lessons learned can be transferred from one to the other.

3 Standard Conformance in Linked Open Data and schema.org Microdata

schema.org provides a formal schema definition for the classes and properties to be used for annotating data. This allows for analyzing the conformance to standard, schema, and best practices, as it is has been done in various places for Linked Open Data as well [3, 7]. In [4], we compare the conformance of schema.org to that in Linked Open Data, based on a corpus of RDF extracted from Microdata in the Web Data Commons project[11]. That corpus comprises data from 398,542 pay-level domains (PLDs), with a total of 6.4 billion triples.

The analysis covers the following aspects:

- Usage of wrong namespaces, such as `http://shema.org/`
- Usage of undefined types
- Usage of undefined properties
- Confusion of datatype properties and object properties
- Datatype range violations (e.g., using a number instead of a date)

[9] `http://schema.org/sameAs`
[10] `http://webdatacommons.org/structureddata/2014-12/stats/stats.html`
[11] `http://webdatacommons.org/`

- Property domain violations (i.e., using properties with subjects of a class not contained in the domain definition)
- Object property range violations (i.e., using properties with objects of a class not contained in the range definition)

By comparing the numbers generated from the RDF corpus to those in similar works conducted on Linked Open Data, we could identify a few interesting differences:

- The usage of undefined elements (i.e., types and properties) is less frequent for Microdata than for LOD. For Microdata, 5.6% resp. 9.7% of the documents use undefined types and properties, as opposed to 38.8% and 72.4% of all documents in LOD.
- The confusion of datatype and object properties in Microdata is much larger. In our corpus, 24.35% of all documents use object properties with a literal object, compared to only 8% in LOD.
- Datatype ranges are violated more than twice as often in Microdata than in LOD (12.1% vs. 4.6%). In both cases, date formats are the most frequent problem.
- Domain violations and object property range violations occur slightly more often in Microdata (3.2% of all documents) than in LOD (2.4% of all documents).

Generally, we can see that in absolute numbers, Microdata has a surprisingly high conformance to the schema. The only deviation is the datatype and object property confusion, which can be partly attributed to the way triples are generated from the annotated HTML code. If an object property is used without any subordinate elements, a triple is extracted which contains the subsequent text as a string literal. In [6], it has been argued that parsing the contents into a blank node might be the better option here. We have shown in [4] that this strategy is feasible, and that in many cases, it is even possible to assign a meaningful type to the new blank node.

There are quite a few possible reasons for the quality of schema.org Microdata often being higher than that of LOD. First, there is a direct economic incentive of providing correct Microdata (as it leads to better visibility in search engine results). Second, schema.org is well-documented, with lots of ready-to-use and easy-to-adapt examples on the documenting web pages. Third, content management systems (CMS), such as *Drupal* have adopted schema.org[12] and, with millions of installations, serve as multipliers. Finally, schema.org continuously evolves, taking up users' suggestions, which may also lead to "misused" constructs becoming officially allowed in later releases.

4 Co-Evolution of the schema.org A-box and T-box

To quantify on these hypotheses, we have started a diachronic analysis of deployed schema.org Microdata. To that end, we have taken three snapshots of Mi-

[12] `https://www.drupal.org/project/schemaorg`

crodata, i.e., from 2012, 2013, and 2014, and look at the corresponding schema definitions which were valid at the respective point in time. With that data collection, we are able to analyze both

- The overall convergence (or divergence) of the data, i.e., whether instances of a given class are described more uniformly over time
- Top-down (i.e., schema first) effects, such as the adoption rate of new features in the schema.org standard definition, or the adoption rate of deprecations
- Bottom-up (i.e., data first) effects, such as standard elements being introduced *after* they have been used "inofficially"

For measuring convergence, we use the set of properties which are defined for an instance of a class as a bit vector, and compute the heterogeneity of each class as the normalized entropy rate across all the instances of the class. An increase in the entropy rate reflects a growing heterogeneity, while a decrease reflects a growing homogeneity.

Globally, the entropy drastically drops, so that we can diagnose a strong homogenization of the data. Looking at class-specific differences, we can see that the adoption of schema.org by content management systems (CMS) such as Drupal has lead to an increase of homogeneity (e.g., for classes like `Website` or `Blog`), as well as classes promoted by Google Rich Snippets[13], which lead to better search engine visibility, such as `Product` and `Offer`, and are also extensively documented with ready-to-use examples.

For top-down processes, we compared the usage of classes and properties before and after they were officially included in the schema.org standard. We found that new classes and properties are often adopted very slowly. There are even domains covered by schema.org for which no deployed data can be found at all, such as the medical domain, where a larger vocabulary was bulk-integrated into schema.org.[14] Deprecations, on the other hand, quickly lead to elements in deployed data being replaced by the newly recommended versions.

For bottom-up processes, we also compare the usage of classes and properties before and after their official announcement. We can observe a mild influence on new classes and properties (i.e., they are occasionally used before becoming official). This is particularly visible in *data vocabulary*, the deprecated predecessor of schema.org, still being the second most deployed Microdata vocabulary [5].

There is, however, a strong influence on domains and ranges of existing properties: here, properties are often used in a different context than intended, and this is likely to be reflected in later versions of the standard.

In addition to simply using undefined classes and properties, there is an official mechanism in schema.org, i.e., the extension mechanism.[15] This mechanism allows users to create subclasses and subproperties of existing classes and properties on the fly. Overall, this mechanism is only rarely used, without a measurable

[13] `https://developers.google.com/structured-data/rich-snippets/`
[14] `http://blog.schema.org/2012/06/health-and-medical-vocabulary-for.html`
[15] `http://schema.org/docs/extension.html`

impact of classes and properties used as extensions first becoming part of the standard later.

It is particularly noteworthy that schema.org is in a state of constant evolution. In the past three years, more than 25 revisions have been published. Together with the fact that both bottom-up and top-down processes can be observed, where the deployed data influences the schema, we can see that there is a co-evolution of data (A-box) and schema (T-box) in schema.org, which is rarely observed for Linked Open Data. For comparison, *FOAF*[16], the most widely deployed LOD vocabulary [7], has undergone only six revisions within the past eight years.

5 Conclusion and Outlook

In this paper, we have contrasted the usage of schema.org Microdata and Linked Open Data. We have looked at standard conformance for Microdata, taking from a synchronic and a diachronic perspective. We have identified several drivers of Microdata adoption:

Business Incentive Whenever there is a direct incentive to use Microdata, such as a better listing in search engine results, we can observe that the schema is followed more strictly.

Availability of Documentation Ready-to-adapt examples increase standard conformance.

Implementation in Widely Deployed Platforms Content-management systems like Drupal use schema.org, which leads to a large-scale usage (sometimes even unconscious to the website owner) and, at the same time, a larger homogeneity of the provided data.

Standard Flexibility If the standard is violated in the same way at larger scale, this may hint at a shortcoming in the standard. schema.org frequently adopts to violations by either declaring them valid, or by offering solutions for the gaps that are filled by the violations.

The first two findings are also supported by the rare adoption of schema.org's extension mechanism. There is hardly a business incentive to define a class via the extension mechanism (if it manages to parse it correctly, a data consumer is likely to treat it exactly the same as the defined super-type), and, in contrast to the rest of the schema, the extension mechanism is only described in a rather far-off section of the schema.org documentation pages.

For Linked Open Data, we can state that things are slightly different. There are no major drivers like the big search engine companies promoting schema.org, which have lead to a dramatic increase of available schema.org Microdata (the adoption of schema.org has grown by roughly a factor of 10 during the past two years). Documentation is often scarce and/or at a deep technical level, and data is provided by technology evangelists rather than commercial providers.

[16] `http://www.foaf-project.org/`

Last, schema flexibility is not as strongly observable as for schema.org, as the comparison with FOAF shows.

In summary, although Microdata and Linked Open Data have some essential differences, they are similar enough to make a comparison feasible and reasonable. Some of the factors identified driving the quick adoption of schema.org Microdata are also interesting findings which could be adopted to further push the adoption of Linked Open Data.

Acknowledgements

The author would like to thank Robert Meusel and Christian Bizer for their valuable ideas, input and analyses, part of which are reflected in this paper.

References

1. Bizer, C., Eckert, K., Meusel, R., Mühleisen, H., Schuhmacher, M., Völker, J.: Deployment of rdfa, microdata, and microformats on the web–a quantitative analysis. In: The Semantic Web–ISWC 2013, pp. 17–32. Springer (2013)
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)
3. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the pedantic web. In: Linked Data on the Web (2010)
4. Meusel, R., Paulheim, H.: Heuristics for fixing errors in deployed schema.org microdata. In: Extended Semantic Web Conference (2015), to appear
5. Meusel, R., Petrovski, P., Bizer, C.: The webdatacommons microdata, rdfa and microformat dataset series. In: ISWC (2014)
6. Patel-Schneider, P.F.: Analyzing Schema.org. In: International Semantic Web Conference (2014)
7. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the linked data best practices in different topical domains. In: International Semantic Web Conference (2014)