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

Heiko Paulheim, Data and Web Science Group, University of Mannheim
Synopsis

• Microdata/schema.org vs. Linked Open Data
  – Microdata/schema.org in a Nutshell
  – Commonalities & Differences
• schema.org adoption and conformance
• Co-evolution of schema.org and deployed data
• Conclusion and Take-aways

Disclaimer:
I do not speak for the schema.org community. All views expressed in this talk are my own.
Microdata in a Nutshell

• Adding structured information to web pages
  – By marking up contents
  – Arbitrary vocabularies are possible
• Similar to RDFa

```html
<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>
```
Microdata in a Nutshell

- Markup can be extracted to RDF

  – See W3C Interest Group Note: Microdata to RDF [1]

```
<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 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/addressCounty> "Germany" .
```

schema.org in a Nutshell

• Microdata can be used with *any* vocabulary
  – Practically, only schema.org is deployed on a large scale
  – Plus its historical predecessor, data-vocabulary.org

• schema.org
  – Vocabulary for marking up web content
  – Promoted and used by major search engines
  – Google, Bing, Yahoo!, and Yandex
  – Can be used with Microdata and RDFa
• ...but RDFa is hardly used
schema.org in a Nutshell

- schema.org (as of May 2015, release 2.0)
  - 675 classes
  - 965 properties
schema.org in a Nutshell

• schema.org has incorporated some popular vocabularies, including
  – Good Relations (2012)
  – W3C BibExtend (2014)
  – Automotive Ontology (2015)
So...

I'M HERE FOR LINKED OPEN DATA

SO WHY SHOULD I CARE?
Microdata/schema.org vs. LOD

• Commonalities
  – Both encode machine-interpretable knowledge
  – Schema.org uses a standard vocabulary
  – Both can be encoded as RDF
Microdata/schema.org vs. LOD

• Differences
  – Microdata is embedded in the DOM tree
  – i.e., the resulting RDF is always a set of trees
  – not a general directed graph
  – no cycles, no reification
  – Microdata uses only blank nodes and literals
Microdata/schema.org vs. LOD

• Linked Data Principles (TimBL 2006)
  – Use URIs as names for things
  – Use HTTP URIs that can be looked up
  – When someone looks up a HTTP URI, provide useful information using a standard
  – Include links to other URIs

MD2RDF creates blank nodes
Blank nodes cannot be looked up
HTML5+MD is a standard
This is possible with schema.org/sameas
Microdata/schema.org vs. LOD

- Linked Data Principles (TimBL 2006)
  - Use URIs as names for things
  - Use HTTP URIs that can be looked up
  - When someone looks up a HTTP URI, provide useful information using a standard

```
<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>
```

[MD2RDF creates blank nodes]
[Blank nodes cannot be looked up]
[HTML5+MD is a standard]
Microdata/schema.org vs. LOD

• Linked Data Principles (TimBL 2006)
  – Use URIs as names for things
  – Use HTTP URIs that can be looked up
  – When someone looks up a HTTP URI, provide useful information using a standard
  – Include links to other URIs

This is possible with schema.org/sameas

• Linkage within schema.org Microdata:
  – Only 0.02% of all data providers use schema.
Microdata/schema.org vs. LOD

• Linked Data Principles (TimBL 2006)
  – Use URIs as names for things
  – Use HTTP URIs that can be looked up
  – When someone looks up a HTTP URI, provide useful information using a standard
  – Include links to other URIs

This is possible with schema.org/sameas

• Linkage within schema.org Microdata:
  – Only 0.02% of all data providers use schema.
Microdata/schema.org vs. LOD

• Linked Data Principles (TimBL 2006)
  – Use URIs as names for things
  – Use HTTP URIs that can be looked up
  – When someone looks up a HTTP URI, provide useful information using a standard
  – Include links to other URIs

  This is possible with schema.org/sameas

• Linkage within schema.org Microdata:
  – Only 0.02% of all data providers use schema.
Microdata/schema.org vs. LOD

• Five Star Scheme (TimBL 2010)
  * Available on the web with an open license
  ** Available as machine-readable, structured data
  *** as (**), using a non-proprietary format
  **** plus: using open standards by the W3C
  ***** plus: links to other datasets

• What's the license of web data?
Analyzing schema.org Microdata

• Web Data Commons project
  – Extracts different sorts of data sets from public web crawl
  – RDFa, Microdata, Microformats
  – Graph structure of the Web
  – Web Content Tables
• http://webdatacommons.org/

We use that corpus for quantitative analysis
Microdata/schema.org vs. LOD

LOD Cloud 2012

LOD Cloud 2014

Number of PLDs Deploying the three Major Markup Formats

Number of Triples marked up by the three Major Markup Formats
...why we should care

• MD/schema.org and LOD have a lot in common
• schema.org is a success story
  – Hundreds of thousands of adopters
• schema.org is large-scale
  – Billions of triples (using the same schema!)
schema.org vs. LOD: Topical Domains

• Domains by number of datasets in Linked Open Data
  – As of 2014
  – Classified based on data provider tags
  – More than half of the datasets is social web (mostly FOAF files)
Main topics of schema.org:
– Meta information on web page content (web page, blog…)
– Business data (products, offers, …)
– Contact data (businesses, persons, …)
– (Product) reviews and ratings
• and a massive long tail
schema.org vs. LOD: Topical Domains

- schema.org
  - High increases in some areas
schema.org Compliance vs. LOD Compliance

• A closer look:
  – How is schema.org actually used?
  – Where is the schema violated?
• Comparison with LOD
  – Hogan et al. (2010): Weaving the Pedantic Web
  – Schmachtenberg et al. (2014): Adoption of the Linked Data Best Practices in Different Topical Domains
schema.org Compliance vs. LOD Compliance

• Compliance dimensions analyzed
  – Usage of wrong namespaces, such as http://schema.org/
  – Usage of undefined types
  – Usage of undefined properties
  – Confusion of datatype properties and object properties
  – Property domain violations
  – Object property range violations
  – Datatype range violations
Building a Corpus of schema.org Microdata

• Starting point: all pages in the CommonCrawl that contain Microdata
• What could be (meant to be) schema.org?
  – Everything that contains “schema.org” as a substring in a namespace
  – Everything that contains URIs where the protocol and authority is similar to http://schema.org/ (within an EditDistance of 1)
  – Noise filtering: removing all namespaces that occur only once
Building a Corpus of schema.org Microdata

• More than 98% of the preselected pages use a correct namespace
• Frequent namespace variations
  – http://www.schema.org/
  – https://schema.org
  – http://schema.org
  – ...

Common Crawl
Building a Corpus of schema.org Microdata

• Final corpus of schema.org Microdata
  – 6.4 Billion triples
  – extracted from 217,018,636 URLs
  – belonging to 398,542 PLDs
• This corresponds to ~86% of the overall Microdata collection
schema.org Compliance vs. LOD Compliance

• Usage of undefined types (6.07% of all PLDs)
• Typical causes:
  – Misspellings: http://schema.org/Store
  – Miscapitalization: http://schema.org/localbusiness

• Comparison:
  – 5.8% of all Microdata documents
  – 38.8% of all LOD documents
  (Hogan et al., 2010)
schema.org Compliance vs. LOD Compliance

- Usage of undefined properties (3.9% of all PLDs)
- Main causes:
  - Miscapitalization: http://schema.org/contentURL
  - Close, but miss
    - http://schema.org/currency
    - http://schema.org/fax
  - Made up
    - http://schema.org/blogId
    - http://schema.org/postId
- Comparison:
  - 9.7% of all Microdata documents
  - 72.4% of all LOD documents
(Hogan et al., 2010)
Using Object Properties as Data Properties (56.6% of all PLDs!)
i.e., using an object property with a string value

Typical properties:
- http://schema.org/addresscountry
- http://schema.org/manufacturer
- http://schema.org/author
- http://schema.org/brand

Comparison:
- 24.35% of all Microdata documents
- 8% of all LOD documents

(Hogan et al., 2010)
schema.org Compliance vs. LOD Compliance

• Using Data Properties as Object Properties
  – i.e., using a data property with a complex object
• Neglectable on Microdata (0.2% of all PLDs)
• Comparison:
  – 0.6% of all Microdata documents
  – 2.2% of all LOD documents
(Hogan et al., 2010)
schema.org Compliance vs. LOD Compliance

• Property Domain Violations (4.0% of all PLDs)
  – i.e., using a property with a subject not included in its domain
• Typical problem: shortcuts, e.g.
  – price on used on Product
  – streetAddress used on LocalBusiness
  – …
• Difficult to compare directly
  – Semantics are different
  – schema.org: domain list is exhaustive
  – LOD: open world assumption
schema.org Compliance vs. LOD Compliance

• Datatype range violations (9.6% of all PLDs)
  – i.e., using a data property with an “incompatible” literal
  – we use an optimistic parser for URLs, numbers and dates
• Top 20 problems
  – 13/20 are dates
  – Three are URLs
  – Two each are numbers and times
• Comparison:
  – 12.06% of all Microdata documents
  – 4.6% of all LOD documents (Hogan et al., 2010)
  – Dates are the main problem in both cases
• Object Property range violations (8.6% of all PLDs)
  – i.e., using an Object Property with an object type outside its range
• Typical problem:
  – mainContentOfPage with http://schema.org/Blog
  – ...instead of http://schema.org/WebPageElement
  – Hints at a missing hierarchy relation?
• Difficult to compare directly
  – Semantics are different
  – schema.org: domain list is exhaustive
  – LOD: open world assumption
schema.org Compliance vs. LOD Compliance

• Microdata compliance to schema.org is surprisingly high
  – Providers are often not technology evangelists
  – (unlike for LOD)
• Most often higher than for LOD
  – Notable exception: confusion of Datatype and Object Properties
• Many of the problems presented can be fixed
  – On the fly on the consumer side
  – With a set of simple heuristics

• Advertisement:
  – Robert Meusel, Heiko Paulheim:
  – Heuristics for Fixing Common Errors
  – in Deployed schema.org Microdata
  – Thursday, 2pm, Room Emerald 1
Digging for Reasons

• So, Microdata is often more schema compliant
• But why? Some hypotheses…
  – Availability of documentation
  – Tool support
  – Business incentive
  – Schema flexibility
• Can we confirm/reject those from looking at the data?
A Diachronic Perspective

• Versions of schema.org are archived over time
  – Plus: there are several crawl releases per year
  – i.e., we can look at change over time
• If we look at both schema and deployed data, we may observe
  – Adoption rates of schema changes
  – Data-first changes to the schema
  – Convergence or divergence of deployed data
A Diachronic Perspective

• Three releases of WDC Microdata corpus
  – 2012, 2013, and 2014
• Versions of schema.org that were valid
  – At the beginning of the crawl
  – At the end of the crawl
Overall Convergence

• Measuring convergence
  – i.e., homogeneity of descriptions of classes
  – Example: two instances of LocalBusiness

```
_:1
  \ s:LocalBusiness
    \  \ s:PostalAddress
      \    \  "Birmingham"
        \  "Main Street 24"

_:2


  \ s:LocalBusiness
    \  \ s:PostalAddress
      \    \  "Liverpool"
        \  "Church Street 1"

```
Overall Convergence

• Recap
  – RDF from Microdata is a set of trees
  – i.e., we can enumerate all paths to leaf nodes (omitting literals)
• Example:
  – rdf:type s:LocalBusiness,
    s:address rdf:type s:PostalAddress,
    s:address s:addressLocality,
    s:address s:streetAddress

```
_:1
  ^
  |    s:LocalBusiness
  |    ^
  |    |    s:PostalAddress
  |    |    ^
  |    |    |    "Liverpool"
  |    |    ^
  |    |    |    Church Street 1
```

...
Overall Convergence

• Using all paths, we can compute the entropy for each class as

\[ H = \sum_{i=1}^{n} -p(path_i)\log(p(path_i)) \]

• A low entropy refers to a high homogeneity
• We normalize both by maximum entropy and the total number of paths
  – i.e., we use normalized entropy rate as a measure for homogeneity
Overall Convergence

• Observations
  – Overall entropy decreases over time

• Classes with high convergence rates
  – WebSite, Blog, …
  – Hotel, Restaurant, …
  – Product, Offer, …
  – Rating, Review

Influence of CMS adoption
Yellow pages
Google Rich Snippets
...all of the above
Top-down Adoption

• How fast are changes in the schema adopted?
  – New classes/properties
  – Deprecations
  – Domain/range changes

• Measuring adoption: challenges
  – Different crawls
  – Overall growth of deployed Microdata

• Measure: normalized usage increase \( (nui) \) from \( i \) to \( j \):
  – \( nui(s) > 1.05 \): usage of schema element \( s \) has increased significantly
  – \( nui(s) < 0.95 \): usage of schema element \( s \) has decreased significantly

\[
nui_{i,j}(s) := \frac{\#PLD_i(s)}{\#PLD_j(s) + 1} / \frac{\#PLD_i}{\#PLD_j} \quad (i > j)
\]
Top-down Adoption

• Adoption of new classes and properties
  – Almost half of all introduced classes are *never* used!
  – Similar for new properties

• Reasons
  – Bulk-addition of vocabularies
  • not every term is equally needed
  • e.g., medical vocabulary
  – Blind spot of our approach
  • some terms are mainly for *e-mail* markup
  • e.g., Actions
Top-down Adoption

• Main domains of positive adoption
  – Meta data for web content (schema.org/Website has the highest NUI)
  – Broadcasting (e.g., TV Episodes)
  – Questions & Answers
  – Postal addresses

• Classes featured in Google Rich Snippets
  – Still growth on high level (tens of thousands of PLDs)
  – But NUI<0.95

Influence of CMS adoption
Collaboration with BBC and EBU
Q&A Pages, such as Stackoverflow
Yellow Pages Search Engine Listings
Top-down Adoption

• Adoption of domain/range changes
  – Again: rather low overall adoption
• Adopted well for
  – Products (height, width, itemCondition, …)
  – Broadcasting domain (episode, season, actor, …)
Top-down Adoption

• Adoption of deprecations
  – Works well (29 out of 32 have a significantly low NUI)
• Exceptions
  – s:map (← s:hasMap)
  – s:maps (← s:hasMap)

For Google Maps
(lots of outdated tutorials)
Bottom-up Evolution

• Martin Luther
  – Started the protestant church
  – A success story, too (like schema.org)
  – (i.e., 800 million adopters worldwide)

• Famous quote:
  – “Man muss [...] dem gemeinen Mann aufs Maul schauen”
  – (roughly: “You have to listen to the way the common man really speaks.”)

Disclaimer: I do not speak for the protestant church.

Martin Luther, 1483-1546
Bottom-up Evolution

• Are new features in the schema first used “inofficially”?  
  – New classes/properties  
  – Domain/range changes

• Instrument for measurement: ROC curves  
  – True positives mapped against false positives  
  – tp: elements used before  
  – fp: elements not used before  
  – Ranking by #PLDs
Bottom-up Evolution

• There are some mild influences observable
  – Stronger for domain/range changes
• especially range changes
  – Weaker for new classes/properties
Bottom-up Evolution

• Extension mechanism
  – Allows for user-defined classes/properties
  – Those become subclasses implicitly

• Analysis over time
  – No measurable impact on standard evolution
  – “Inofficial” use is likelier than use of extension mechanism
Key Adoption Drivers

• Search Engine Optimization
  – Web site providers want to be high in Google rankings
  – Direct business incentive!
• Tool adoption
  – Major CMSs use schema.org
• Standard Agility
  – schema.org: 25 revisions in last three years
  – cf. FOAF: six revisions in last eight years
What Does that Mean for LOD?

• Schema.org has been a success story
• Mainly because of
  – Business incentive
  – Documentation
  – Platform adoption (marketing!)
  – Standard agility

I'M HERE FOR LINKED OPEN DATA
SO WHY SHOULD I CARE?
What Does that Mean for LOD?

• Business Incentive for Data Providers
  – not: the cool thing I can do if you provide your data as LOD
  – but: what's in for you if you provide your data as LOD

• Documentation
  – schema.org has copy-and-paste snippets for HTML5
  – what's the effort of setting up a LOD endpoint?

• Platform Adoption
  – Think: What if MySQL had an out-of-the-box LOD endpoint?

• Standard Agility
  – Observe how your schemas are “abused”
  – ...and don't blame the abusers
Wrap-up

• We have looked at the deployment of schema.org
  – From a synchronic and a diachronic perspective
• From looking at the data, we can observe
  – Driving factors for adoption
  – Patterns in standard evolution
  – Influence of data consumers, documentation, tool adoption
• ...and identify factors that lead to the wide adoption
• LOD and schema.org/MD have a lot in common
  – so we may take away some lessons for boosting the adoption of LOD...
Acknowledgements

• Colleagues at Data&Web Science Group who contributed to this talk
  – Robert Meusel (Number crunching)
  – Christian Bizer (Fruitful discussions)
• ...and the Common Crawl Foundation
  – Who makes such an analysis possible
• Want to get your hands dirty with data?
• Linked Data for Information Extraction (LD4IE) challenge 2015
  – Use schema.org Microdata
  – Train a web-scale IE system
  – Present at LD4IE workshop (co-located with ISWC 2015)
  – Win a prize sponsored by Springer!
• Details
  – http://oak.dcs.shef.ac.uk/ld4ie2015
Thank you! Questions?
What the Adoption of schema.org Tells about Linked Open Data

Heiko Paulheim, Data and Web Science Group