This is a position paper from schema.org for the W3C Open Data on the Web Workshop.

Authors: Dan Brickley (Google), R.V. Guha (Google), Steve Macbeth (Microsoft), Peter Mika (Yahoo), Alexander Shubin (Yandex)

Schema.org provides a collection of schemas typically used as annotations on HTML tags. Webmasters can use the types and properties defined at schema.org to markup their pages in ways recognized by major search providers. This paper offers a few informal observations about how schema.org fits into the wider 'open data' Web community.

Schema.org recap

Schema.org is designed for extremely mainstream, mass-market adoption. This required an overriding emphasis on simplicity for publishers. This affects our technical strategy in several ways:

- It is a relatively large vocabulary, compared to typical RDF vocabularies.
    - We don't want publishers to have to deal with dozens of independent vocabularies whose interconnections are undocumented.
    - Schema.org brings together several independently defined vocabularies (e.g. rNews, LRMI, Good Relations); credits for these different sources are cited once from the relevant schema.org type pages, rather than in every page using the vocabulary.
- In this environment, it is very difficult to be purist. For example, although everyone prefers structured data to make a clean distinctions between IDs for Web pages versus for the things they describe, we take a 'something is better than nothing' view. Schema.org data can be messy. Our starting point is the existing Web, and that has guided our technology choices. Aside from such pragmatic differences of emphasis, schema.org can be considered an approach to 'Linked Data'.
- Schema.org's primary usage is for annotating existing Web content. This requires our vocabulary to often be flatter and less normalized than a purely database-oriented approach might be.
- Schema.org is based on W3C's RDF model for structured data. The datamodel is essentially RDF triples, the schema is published using an extended subset of RDFS, and it can be exchanged using any RDF syntax, alongside Microdata. Microdata itself was based on RDFa 1.0, but simplified for mainstream Web publisher use. Microdata's improvements to the RDFa 1.0 syntax have been adopted in W3C's RDFa 1.1 as 'RDFa Lite', and allow publishers to use schema.org easily alongside other RDF vocabularies.
- Although schema.org contains hundreds of terms, there are many areas it cannot cover in detail. Ongoing additions are discussed in the W3C Web Schemas community. We have also articulated the notion of an 'external enumeration' to make clear how schema.org can be combined with larger vocabularies and datasets from elsewhere (e.g. Wikipedia/Wikidata, SKOS, Freebase).

- Schema.org emphasises simplicity for publishers, and - following the microformats and RDFa tradition - encourages higher quality data by having multiple independent consuming applications and by emphasising annotation of human-facing pages rather than hidden (and potentially neglected) machine-only feeds.

Data on the Web - beyond 'Graphs'

The notion of a 'graph' model for structured data is enjoying some popularity. Talk of 'social graphs', 'interest graphs', alongside the original 'Web graph' emphasise the usefulness of this simple entity-relationship view of data. However, it has limits.

Schema.org's core datamodel is very much in the 'graph' tradition, and allows all kinds of domain-oriented, descriptive graphs to be overlaid on top of the classic hypertext "Web graph".

Schema.org defines a set of types, and a set of properties (i.e. relationship types). This is a simple and powerful formalism, making it possible for schema.org to look beyond its current HTML focus at a more general role in data interoperability. The underlying graph model has some conceptual elegance (entities plus their properties, types and relationships), allowing the basic idea to be easily communicated. It is also a model not tied explicitly to specific file formats (XML, JSON(-LD), ...). This last aspect is both a strength and a weakness; e.g. see earlier debates around XML and schema languages.

Having said all this, we should be very clear that entity-relationship graphs are *not* always the most appropriate data representation. It is important to state clearly that no single data format or abstract data model fully addresses universal needs. While abstract entity-relationship graphs such as RDF's offer some protection against 'fashions' for more concrete formats (XML, JSON, … whatever comes next), this comes at some cost. The extra layer of indirection involved in defining these more abstract schemas can be a barrier to entry, making an emphasis on simplicity and mass-market adoption all the more critical. These tradeoffs also requires pragmatism: sometimes structured data in entity-centric graph form is only part of the story.

- Much data is tabular and numeric.
- Other datasets are in custom application-oriented forms.
- Often datasets need per fact annotations and footnotes, or other qualifiers.
- Topic/Subject description (e.g. bibliographic) often has a subtler notion of 'aboutness' than "about this entity".
- Often publishers and consumers of data want larger 'units of information' than simple factual triples, e.g. to preserve the packaging/integrity of some set of records.

To some extent, the graph datamodel can be adapted (*stretched...*) to deal with these considerations. For example:

- The W3C 'data cube' work allows tabular/numeric datasets to be described in RDF.

- W3C SKOS allows topics ('concepts') to be described as first class entities.
- RDF reification and the Named Graph mechanism from SPARQL provide (sometimes awkward) machinery for allowing pieces of data to be described.
- W3C's OWL language has various constructs that are encoded as RDF triples, although they make more sense when viewed at a more abstract level.

However it is important to be pluralistic. RDF-style graphs are not always the most appropriate representation for all kind of data. When an application of RDF slips into heavy use of a single type (e.g. 'Record'), this is often a hint that the graph representation is serving only as a conduit. Although RDF-shaped graphs are expressive, sometimes it is best to use RDF for its original metadata role. Graph-structured data can very usefully describe data in a variety of forms, linking it with other items, topics, its history, provenance and social context. The actual data files might be in CSV, JSON, XML or other form; nevertheless, a graph-based description can be used to link the dataset into a wider Web of structured data.

For this reason we are adding some basic abilities to schema.org for *describing* datasets.

Future Work?

Perhaps edge-labeled graphs can play a similar role for structured data in the Web as HTML has played for human-facing documents? The Web is a thriving mix of many document formats, yet is held together in large part by one: HTML. Simple graph-structured (meta)data has the potential to serve as a general purpose 'table of contents' for a similar Web of open data formats and services. Schema.org in turn has the potential to help bootstrap this by providing a general purpose shared utility vocabulary that addresses many common use cases, while remaining extensible for more focussed applications.

There are many aspects of dataset description and discovery that are not yet well addressed by schema.org vocabulary. As schema.org begin to add basic terms such as 'Dataset', 'DataDownload', 'DataCatalog' this W3C Workshop provides a forum in which other aspects of dataset description can be discussed. It is likely that schema.org will go on to explore richer ways of 'looking inside' a dataset, e.g. by annotating how to expand dataset records into triples/graphs, or link per-dataset identifiers with externally identifiable entities and topics. However it may be equally important to anchor datasets in their larger social context, associating them with institutions and processes, researchers and groups, information about provenance, quality and workflow, scholarly publications, conference presentations, funding, peer review, visualizations, community and commentary. The cross-domain nature of schema.org can help here, as it provides terminology across many of these areas. W3C's Web Schemas forum provides an open community environment within which suggestions for improved schemas for datasets can be discussed; responses to this paper are welcomed there.

There are many aspects to 'open data on the Web' not touched upon here. Specifically this paper does not address W3C's ongoing work around HTML, or the family relationships amongst

e.g. different flavours of RDFa, Microformats and Microdata. This is not to say they're unimportant, and there are many opportunities ahead in the browser/HTML area (e.g. around improvements to Web forms, representations of potential and actual 'Actions', etc.). The focus of this paper is to emphasise that even the introduction of a simple notion of a 'Dataset', as a larger package of information than a triple or edge in a graph, is a useful contribution for graph-based metadata. Rather than trying to fit all kinds of data directly into the edge-labeled 'graph' datamodel, we can instead use such graphs as metadata to describe various kinds of dataset, giving hints not only to their factual contents, but to their wider social context too.