2014

RDFa and Microdata

Bethany Wetherill
San Jose State University, bethanywetherill@gmail.com

Running head: RDFa and Microdata

RDFa and Microdata: Is One Better Than The Other?

Bethany Wetherill

A Research Proposal Submitted
in Partial Fulfillment of the Requirements
for LIBR 281 - Metadata

San Jose State University
2014

Abstract

This project seeks to explore and observe differences in RDFa and microdata and their ability to retain proper schematization and syntax when converted back to RDF/XML. Online conversion tools were used to transpose existing RDF/XML files from online data dumps to RDFa and microdata, and then back to RDF/XML, offering some insights into RDFa and microdata's capabilities, as well as a taste of what may happen in the future if major search engines decide to move away from microdata and developers need to convert to a different semantic markup language. Multiple online converters were employed in the conversion process in an attempt to identify and isolate the effects of potential programming errors. Ultimately, there was no major difference in data integrity between RDFa to RDF/XML and microdata to RDF/XML. Larger differences can be ascribed to the conversion tools being used rather than the formats themselves, as well as the quality the original RDF document.

RDFa and Microdata: Is One Better Than The Other?

The resource description framework (RDF) was developed by the World Wide Web Consortium (W3C) a standard for organizing the semantic web. One of the most commonly cited issues with RDF is that many times developers don't think in terms of interoperability, and thus, as that is its primary selling feature, RDF is slow to be adopted (Alam, Khan, & Thuraisingham, 2011; Davis, 2011). This is less an issue when looking at the web in general, as many sites are designed with search engine optimization in mind and are thus keen on adhering to whatever standards make content accessible to search engines. The complexity of RDF, however, led to slow adoption because it was not feasible for the average web developer to implement. To solve this issue, RDFa (RDF in attributes) was introduced to allow RDF expressions within HTML, XHTML, and XML documents. RDFa, like RDF, is a product of the W3C, and was their answer to accessibility and usability by web developers.

RDF and even RDFa (despite being designed for usability) require some knowledge of computer science and/or metadata and study of how the fairly complex systems work, which can be intimidating at best and in many cases a barrier. Microdata was developed by the Web Hypertext Application Technology Working Group (WHATWG) as an alternative model for including metadata in HTML documents. Microdata aimed to be simpler yet in order to facilitate adoption by as many developers as possible. Schema.org has become the home of all things microdata, as it contains a primer on using microdata and the most set of microdata terms (itemtypes and properties) supported by the major search engines. A major criticism of microdata, however, is that is isn't sophisticated enough to adequately and accurately convey resource metadata.

This project seeks to explore and observe differences in RDFa and microdata and their ability to retain proper schematization and syntax when converted back to RDF/XML. Online conversion tools were used to transpose existing RDF/XML files from online data dumps to RDFa and microdata, and then back to RDF/XML, offering some insights into RDFa and microdata's capabilities, as well as a taste of what may happen in the future if major search engines decide to move away from microdata and developers need to convert to a different semantic markup language.

Literature Review

Literature on RDFa and microdata found in academic journals comprises a small segment of available literature on the subject. Likely to the comparatively slow publication cycles compared to the speed at which RDFa and microdata are developing, peer-reviewed literature consists primarily of topical overviews and projections of future technologies that may build on RDFa and microdata. The bulk of information relating to RDFa and microdata can be found on the W3C site (RDFa) and schema.org, the home of microdata. Additionally, the major search engines and companies involved in development of these frameworks publish information on their websites. Individuals who are involved or were once involved in development teams for these schema blog prolifically, as do web developers, metadata specialists, and those with specialized interests in the semantic web. This literature review will address a breadth of writings on the development, strengths and weaknesses and various contexts, and future of RDFa and microdata.

Both RDFa and microdata contain embedded semantic data. Adida (2011) put forth four critera for embedded semantic information: impendence and extensibility, data shouldn't be duplicated, users should be able to access structured metadata along with the data within their browser, and data/structured data should be self-contained (to facilitate modern copy/past culture). Of these four criteria, the several deem extensibility and self-containment the most important (Adida, B., 2011; Soylu et al, 2012). Soylu notes, “Microformats and eRDF [embedded RDF, a now obsolete predecessor of RDFa and microdata] lack self-containment because it is not possible to re-use eRDF or microformat information without requiring vocabulary specific information.” Of the four formats of embedded semantic data that have been developed, RDFa and microdata have prevailed.

While somewhat similar in appearance, RDFa and microdata are arose from slightly different points of view. Barker (2012) points out that schema.org is “anchored around humans searching the web,” which requires identifying things and describing their characteristics and relationships to other things within the context of the web, giving it much the same purpose as RDFa. However, while RDFa maintains the traditional RDF subject-predicate-object expressions, and like RDF is intent on defining relationships, microdata has a defined set of entities that offer a simpler structure, with slightly less flexibility, designed to spoon-feed search engines and users with information on the web. While RDFa makes use of URIs to clarify relationships (Sporny, 2011; Seadle, 2013), Sodnik et al. offer a succinct explanation of microdata: “The model of microdata consists of groups of properties named items. The items and their properties are presented in the context of the existing elements. An item can be created using the itemscope attribute in any HTML element” (2013, p. 30). Using the attributes "item", "itemprop" and "subject", microdata can theoretically be used to generate RDF triples (Sporny, 2011; Tomberg & Laanpere, 2009).

RDFa was one the earliest of the formats, and is also the most comprehensive, rich, standard. It was part of XHTML 2 and became a W3C standard in 2008 (Lawson, 2012). It is considered somewhat of a gold standard by many data preservation enthusiasts. Seadle captures the "RDF is best" sentiment with his statement that, "RDF enthusiasts have an almost evangelical conviction that it will transform the web and make information vastly more accessible" (2013, p. 184). Even Yahoo's Peter Mika, despite Yahoo's promotion of microdata, admits, "Personally, I would have liked to see RDFa used as the syntax for the basic examples, because I consider it more mature, and a superior standard to microdata in many ways" (2011).

Google announced support for RDFa in 2009, but then in June 2011 the major search engines, including Google, Yahoo, and Bing launched schema.org and announced that they would henceforth be developing for microdata. The complexity of RDFa, it seems, was too much for webmasters. Google fellow Ramanathan Guha noted that in the years that Google had been supporting RDFa, "One of the startling discoveries we made was that the error rate (i.e., webmasters marking up their pages to say X when the really meant to say Y) was about 3 times as much as it was for other formats (which include microformats, sitemaps, Google shopping feeds, etc.)" (Guha, 2011). So while in theory RDFa holds they keys to much possibility, it may not play out as well in reality as in theory.

Schema.org microdata operates at a more general level of specificity than RDFa, making it easier to implement (Bergman, 2011; Hilliker, Wacker, & Nurnberger, 2013; Ronallo, 2012). It is a newer standard, released with HTML 5 (schema.org, 2014; Lawson 2012). Tomberg and Laanpere (2009) describe microdata as, a "hybrid made from microformats, RDFa and HTML 5 syntaxes." While schema.org and microdata are technically separate things, schema.org is run by the major search engines supporting microdata, and therefore maintains control to the extent that many authors reference the two somewhat interchangeably (Barker 2012; Bergman, 2012; Méndez & Greenberg, 2012).

Several authors reference fine balance between simplicity and extensibility/functionality. Jeni Tennison, software engineer and Technical Director at the Open Data Institute, writes, “The three standards previously used for marking up data include microdata, microformats, and RDFa. Schema uses only microdata because the groups found that microdata strikes a balance between the extensibility of RDFa and the simplicity of microformats” (2011). Soylu also references the “tradeoff between simplicity (i.e., usability) and functionality” (Soylu, p. 96).

Schema.org directly addresses the question, “Why microdata? Why not RDFa or microformats?” in their Frequently Asked Questions. Schema.org acknowledges the decision as a pragmatic one, like Bergman simultaneously acknowledging the superior extensibility or RDFa, noting the necessity for simplicity while still offering some extensibility (Schema.org, 2014).

Outside of the evangelists, several sematic web experts find that microdata is not only sufficient for average use, but will boost RDF. Bergman writes, "In our view, RDF and its triple representations in its data model, is the simplest and most expressive means to represent any data or any data relationship. ...But, simple records and simple data need not be encumbered with the complexity of RDF." (Bergman, 2011) He furthers that just because RDF provides a universal model for data exchange does not mean that it necessarily used; the best exchange formats will be the one that is easy to understand and use, and therefore widely used instead of being a theoretical ideal. Because microdata can be mapped back to RDFa, even if not to full effect, it is still a step toward building the semantic web.

Yahoo’s Peter Mika explains, “schema.org covers the core interest of search providers, i.e., the stuff that people search for the most” (2011). He, like Bergman, asserts that, “should be sufficient to describe the various types of things and their attributes as may be found on the

Web" (Bergman, 2013). The focus is on common web searches like addresses, business reviews, recipes, etc. Mika acknowledges that this does not cover complex subject matters such as biotechnology, e-government, etc. In some cases like these, the complexity of RDFa is still necessary. Floty (2013) emphasizes the need for embedded metadata in web documents and asserts a preference for the power afforded by RDFa (something currently necessary for 3D web components, which he lists as spatial, temporal, structural, logical and behavioral) but does not discount the possible future usability of microdata as it continues to be developed.

While some things may seem too large or complex to be effectively defined in microdata, it is important to differentiate between the backend and user-end of data systems and address the issue of what average users may be looking for. So while the biotechnology example stands, Erik Mitchell of ALA Techsource points out that while most library, archive, and museum systems require much larger and in-depth metadata systems on the backend, that the user-ends of these systems, including the OCLC's WorldCat are beginning to use microdata to enhance findability for users (2013).

In the spirit of comprehensiveness and trying to plan ahead, some consider using both RDFa and microdata within a document. While some argue that this is unnecessary and confusing (Bergman, 2011; schema.org, 2014), others consider it the right thing to do. Tennison justifies her use of both, "Publishers like us at legislation.gov.uk who are aiming to share their data to whoever is interested in it (rather than having a particular consumer in mind) are also likely to want to publish in both microdata and RDFa, rather than force potential consumers to adopt a particular processing model, and will therefore need to mix the syntaxes within their pages" (2011).

Current literature leaves little doubt that microdata is the prevailing format for embedding metadata in web documents. Many agree that microdata is the “best” form of embedding metadata, in that it is easy to use and therefore used and used correctly, unlike RDF(a). There is some doubt, however, that it maintainable, that it can be cleanly mapped back to RDF, and can offer the same richness as RDFa. Whittled down versions of RDFa, namely RDFa-lite, are growing congruently along with microdata, aiming for similar levels of simplicity without making so many compromises on extensibility. While currently supporting microdata, Google, Bing, and Yahoo have not ruled out renewing support for RDFa or other formats in the future. Advances are being made in the automatic translation of microdata to RDFa and vice versa, and even microdata to RDF, such that at some point richness and extensibility may not have to be compromised for ease of use and RDFa may prevail.

Methods

This project sought to observe differences in RDFa, and microdata generated based on identical RDF/XML documents. RDF data sources were selected from the W3C data dump list (http://www.w3.org/wiki/DataSetRDFDumps), Datahub (http://datahub.io/organization/lod), and around the web. Individual data dumps were chosen based on ease of navigability of the website (where the file easy to locate and download), availability of discrete files per record (to avoid large, cumbersome documents as some were only available as single files containing billions of records, weighing in at 20MB while still compressed), availability of RDF/XML records as opposed to other RDF syntaxes like N-Triples (which aren’t so easily readable to humans or plain text editors), and for diversity of different types of resources described in the files (ebooks, legislative data, census data, etc.).

This sample aimed to include a diversity of the types of resources often described on the web – documents are various types, people, places, and objects.

The sample included

1. an ebook from Project Gutenberg (http://www.gutenberg.org/wiki/Gutenberg:Feeds);
2. legislative data from the 103 Congress (https://www.govtrack.us/data/rdf/);
3. encycolopedic information on the city of Berlin from Wikipedia via DBpedia (http://wiki.dbpedia.org/);
4. an entry from the Semantic Bible (http://semanticbible.com/cgi/cgi-in-rdf.html); and
5. encyclopedic information about physics from Wikipedia via DBpedia (http://wiki.dbpedia.org/).

To keep the project manageable in size, smaller RDF/XML files were chosen (some available were so large they would not open in a standard text editing program or could not be converted via current online converters). Files over 200 lines were shortened by removing elements closest to the end with well-formedness of the XML in mind. For example, in a list of 6000 bills and resolutions, some may be removed so there are only 60 bills and resolutions; a node like dcterms:subject with all of its child notes might be removed, removing six lines, while leaving its parent node, pgterms:ebook, structurally intact.

Each of the five records was converted to RDFa and microdata respectively, using free, online browser tools, and the back to RDF/XML. To try to mitigate (or at least more easily identify) problems attributable to the converter rather than the change in format/syntax, two different converters were used for each step when possible. Thus, each of the five records was converted to RDFa and microdata formats using both ebusiness.com and appspot, and the back to RDF via appspot and W3C, for a total of 11 records per original RDF/XML file.

A topical analysis was performed using Oxygen's "compare files" tool, comparing original and converted RDF/XML files for possible changed or missing elements. Settings were adjusted to use XML lines algorithm, Algorithm strength very high; preferences were set to ignore whitespaces, attribute order, comments. Observations were noted, and intermediary files were consulted in the case of unexpected oddities in order to better understand their origins and update observations.

Results & Discussion

Overall, there did not seem to be much difference in data integrity between RDFa to RDF/XML and microdata to RDF/XML. Larger differences can be ascribed to the conversion tools being used rather than the formats themselves, and the original RDF document.

In all cases, elements and URIs were generally maintained; meaning the overall information in the document was still there. Often, initial namespace declarations were changed or substituted, which, again has more to do with the conversion tools than the formats themselves. As part of these missing namespace declarations, former top-level elements were replaced with rdf:Description elements while their child elements remained intact (see Appendix A for full breakdown of differences).

This exploratory project, although in no way definitive, generally supports previous assertions that microdata is sufficient for the most common, generalized uses of metadata on the web: describing people, documents, and businesses. It demonstrates that converting metadata from one syntax and back is much like plain text or spoken words from one language to another and back – it typically results in an awkward approximation of the original, but underlying message is often still there. And like language, any imperfections in the original statement/document will be magnified through this process (which is potentially problematic considering earlier cited statistics about how many faulty RDF documents are on the web).

Because of limitations in comparing the exact contents of the files, and the imperfections caused by the conversion processes, it cannot be concluded whether or not this data supports or refutes the assertion that microdata does not perform as well as RDFa for more complex matters like politics. Perhaps a smaller data sample or different technologies would allow for URI-by-URI comparison and more accurate results in a future project. Additionally, more complex data may need to be obtained to further assess the capabilities of RDFa and microdata for complex documents.

Conclusion

Microdata and RDFa both offer reasonable methods for including metadata in webpages to build the semantic web. While some consider RDF to be the "best" framework for the semantic web, RDFa and microdata offer much more useable options, with microdata winning support of the major search engines while taking criticism from some semantic web scholars for being too simple. This project found no discernible difference in the quality or quantity of metadata retained in RDFa and microdata, at least not attributable to the formats themselves, but rather more contingent on the conversion program. While this conversion issue may come into play as web developers try to modernize and change with changing standards, that is a separate issue and one already being addressed by people building conversion tools.

References

Adida, B. (2011). hGRDDL: Bridging Microformats and RDFa. Web Semantics: Science, Services and Agents On The World Wide Web, 6(1). Retrieved from http://www.websemanticsjournal.org/index.php/ps/article/view/134

Alam, A., Khan, L., & Thuraisingham, B. (2011). Geospatial Resource Description Framework (GRDF) and security constructs. *Computer Standards & Interfaces*, *33*(1), 35–41. doi:10.1016/j.csi.2010.01.002

Barker, P. (2012, August 16.). Where does schema.org fit in the (semantic) web? [Web log post] Retrieved from http://blogs.cetis.ac.uk/philb/2012/08/16/where-does-schemaorg-fit-in-the-semantic-web/

Bergman, M. (2011, June 2). Structured Web Gets Massive Boost. [Web log post] Retrieved from http://www.mkbergman.com/962/structured-web-gets-massive-boost/

Davis, Ian (2011, August 18). The real challenge for RDF is yet to come.

Floty, J. (2013). Describing Semantics of 3D Web Content with RDFa, (c), 63–68.

García, R., Hepp, M., and A. R. (2013). EBusiness RDF2Microdata Converter. Retrieved May 13, 2014, from http://www.ebusiness-unibw.org/tools/rdf2microdata/#convert-by-input

Hilliker, R. J., Wacker, M., & Nurnberger, A. L. (2013). Improving Discovery of and Access to Digital Repository Contents Using Semantic Web Standards: Columbia University's Academic Commons. *Journal of Library Metadata*, *13*(2-3), 80–94. doi:10.1080/19386389.2013.826036

Lawson, B. (2012). On the talismanic fight between RDFa and microdata. [Web log post] Retrieved from http://www.brucelawson.co.uk/2012/talismanic-fight-between-rdfa-and-microdata/

Méndez, E., & Greenberg, J. (2012). Linked data for open vocabularies and *HIVE*'s global framework. *El Profesional de La Informacion*, *21*(3), 236–244. doi:10.3145/epi.2012.may.03

Mitchell, E. T. (2013). Building Blocks of Linked Open Data in Libraries. In *Library Linked Data: Research and Adoption* (Vol. 49, pp. 11–26). ALA TechSource.

Ronallo, J. (n.d.). HTML5 Microdata and Schema.org. *The Code4Lib Journal*, (16). Retrieved from http://journal.code4lib.org/articles/6400?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed:+c4lj+(The+Code4Lib+Journal)

Schema.org. (n.d.). FAQ - schema.org. Retrieved April 07, 2014, from http://schema.org/docs/faq.html

Scott, D. (2013). Microdata□: making metadata matter for machines Evergreen International Conference 2013 Structured data□: making metadata matter for machines Evergreen International Conference 2013.

Seadle, M. (2013). Archiving in the networked world: resource description framework. *Library Hi Tech*, *31*(1), 182–188. doi:10.1108/07378831311304010

Sodnik, J., Jekovec, M., Jakus, G., Tomaţič, S. (2013). The Future of Web. *E-Society Journal*, *2*(1), 27–38. doi:10.5603/KP.2013.0151

Soylu, A., & Felix, M. (2012). Ubiquitous web navigation through harvesting embedded semantic data□: A mobile scenario, *19*, 93–109. doi:10.3233/ICA-2012-0393

Sporny, M. (2011, June 25). An Uber-comparison of RDFa, Microdata and Microformats. Retrieved from http://manu.sporny.org/2011/uber-comparison-rdfa-md-uf/

Stolz, A. (2013). RDF Translator (AppSpot). Retrieved May 13, 2014, from http://rdf-translator.appspot.com/

Tennison, J. (2011, August 20). Microdata and RDFa Living Together. [Web log post] Retrieved from http://www.jenitennison.com/blog/node/165

Tomberg, V., & Laanpere, M. (n.d.). RDFa versus Microformats□: Exploring the Potential for Semantic Interoperability of Mash-up Personal Learning Environments, 102–109.

World Wide Web Consortium (W3C). (2010). RDFa Lite 1.1. Retrieved April 07, 2014, from http://www.w3.org/2010/02/rdfa/sources/rdfa-lite/

World Wide Web Consortium (WC3). (2013). Microdata to RDF Distiller. Retrieved May 13, 2014, from http://www.w3.org/2012/pyMicrodata/#distill_by_input+with_options

World Wide Web Consortium (W3C). (2014). RDF 1.1 Primer. Retrieved from http://www.w3.org/TR/rdf11-primer/

Appendix A: Observation Notes

Key:
AppSpot: RDF Translator: http://rdf-translator.appspot.com/
Ebusiness (RDFa): EBusiness RDF2RDFa Converter: http://www.ebusiness-unibw.org/tools/rdf2rdfa/
Ebusiness (Microdata): EBusiness RDF2Microdata Converter: http://www.ebusiness-unibw.org/tools/rdf2microdata
W3C: Microdata to RDF Distiller: http://www.w3.org/2012/pyMicrodata/

Formatting key:
RDFa/AppSpot/Appspot: Document conferted from RDF/XML to RDFa using AppSpot converter, then RDFa to RDF/XML with AppSpot converter

Sample 1: Project Gutenberg

RDFa/AppSpot-AppSpot – elements and URIs were generally maintained. Unsurprisingly, there were a lot of seemingly extraneous attributes and redundancies, but nothing lost. Some elements were out of order. – Dublin Core element organization and top-level elements lost, former <dcelement> children <description> became top level elements. Lost xml:base=http://www.gutenberg.org/ (present in microdata file from ebiz)

RDFa/AppSpot-EBusiness – Lost xml:base=http://www.gutenberg.org/ (present in microdata file from ebiz), Dublin Core element organization and top-level elements lost, former <dcelement> children <description> became top level elements, some iterations of dc:extent missing

Microdata/AppSpot-Appspot: - Lost xml:base=http://www.gutenberg.org/ (present in microdata file from ebiz), Dublin Core element organization and top-level elements lost, former <dcelement> children <description> became top level elements

Microdata/Ebusiness-Appspot: - Lost xml:base=http://www.gutenberg.org/ (present in microdata file from ebiz), Dublin Core element organization and top-level elements lost, former <dcelement> children <description> became top level elements

Microdata/AppSpot-W3C - Namespace http://www.gutenberg.org/2009/pgterms/ maintained; other namespaces renamed, W3C namespace MD substituted for some elements, but Dublin Core mostly intact

Microdata/Ebusiness-W3C - Namespace http://www.gutenberg.org/2009/pgterms/ maintained; other namespaces renamed with generic number scheme, W3C namespace MD substituted for some elements, but Dublin Core mostly intact

Sample 2: Legslative Data from 103 Congress

RDFa/AppSpot-AppSpot - Error from Appspot

RDFa/Ebusiness-appspot - All namespaces from http://www.rdfabout.com (legislative-focused elements) gone, along with FOAF, but rdfabout elements are used in text
<bill:HouseBill> top element gone and nesting levels, but other bill: elements remain intact

Microdata/AppSpot/AppSpot - Same as above

Microdata/AppSpot/W3C - Same as above; some namespaces renamed with generic names/numbers

Microdata/Ebusiness/Aps - Missing rdfabout namespaces, same as above

Microdata/Ebusiness/W3C – Same as above

Example 3: DBPedia Berlin article

RDFa AppSpot/AppSpot - A few missing namespaces, schemas declared inline instead of at top of document, restructuring less noticeable than in previous examples because rdf:Description was already used as a top-level element

RDFa/ Eusiness/ApsSpot - As above

Microdata Ebusiness/W3 - Namespacs renamed or removed, nesting gone awry and things linked via URI instead

Microdata Ebusiness/Appspot - A few missing namespaces, schemas declared inline, restructuring less noticeable because rdf:Description was already used as a top-level element

Microdata AppSpot/W3 - Namespacs renamed or removed, nesting gone awry as above

Micro AppSpot/AppSpot - A few missing namespaces, schemas declared inline, restructuring less noticeable because rdf:Description was already used as a top-level element

Example 4: Semantic Bible

RDFa AppSpot/AppSpot - Interestingly, namespaces had been generated where there weren't any at the top of the document previously. Lost top level element <cgi:Pericope, others out of order and nested under rdf:Description, otherwise cleanest so far

RDFa Ebusiness/AppSpot - As above

Microdata AppSpot/AppSpot - As above

Microdata AppSpot/W3C - Like other W3C things, weird order and renaming of namespaces. This was the only iteration where Pericope element was kept as top level and intact

Mircodata Ebusiness/AppSpot - Add name spaces, nesting lost but similar order, rdf:Desciption at the top in place of Pericope

Microdata Ebusiness/W3C - Like other W3C things, weird order and renaming of namespaces, kept Pericope element as top level and intact

Example 5 – DBPedia Physics

RDFa AppSpot/AppSpot - A few missing namespaces, schemas declared inline instead of at top of document, restructuring less noticeable than in previous examples because rdf:Description was already used as a top-level element

RDFa/ Eusiness/ApsSpot - As above

Microdata Ebusiness/W3 - Namespacs renamed or removed, nesting gone awry and things linked via URI instead

Microdata Ebusiness/Appspot - A few missing namespaces, schemas declared inline, restructuring less noticeable because rdf:Description was already used as a top-level element

Microdata AppSpot/W3 - Namespacs renamed or removed, nesting gone awry as above

Micro AppSpot/AppSpot - A few missing namespaces, schemas declared inline, restructuring less noticeable because rdf:Description was already used as a top-level element