



Linked Data tools: Semantic Web for the masses

by Lisa Goddard
and Gillian Byrne

Abstract

Semantic Web technologies have immense potential to transform the Internet into a distributed reasoning machine that will not only execute extremely precise searches, but will also have the ability to analyze the data it finds to create new knowledge. This paper examines the state of Semantic Web (also known as Linked Data) tools and infrastructure to determine whether semantic technologies are sufficiently mature for non-expert use, and to identify some of the obstacles to global Linked Data implementation.

Contents

[Introduction](#)

[Are Semantic Web technologies ready for production?](#)

[Exposing data as RDF](#)

[Linking RDF entities together](#)

[Non-technical barriers to broad implementation of Web 3.0](#)

[Conclusion](#)

Introduction

When Tim Berners-Lee declared the Semantic Web “open for business” in February 2008 (Miller, 2008) there were some fairly skeptical responses, even from within the Semantic Web community. Critics said that the RDF standard is too complex and difficult to implement, that named entity markup is too labor intensive to be practical, and that creating agreed-upon ontologies to model all of the world’s knowledge was so gargantuan a task as to be impossible.

Despite all of this, 2009 proved to be a bumper year for Linked Data. Both the U.K. and the U.S. governments unveiled public data Web sites amidst promises to radically open up data and promote transparency. The U.K. even declared an ambitious plan to “aim for the majority of government-published information to be reusable, Linked Data by June 2011” [1]. Global media agencies like the British Broadcasting Corporation (BBC) and the *New York Times* (NYT) began to expose their huge stores of data in Resource Description Framework (RDF) and link them to other Semantic Web vocabularies. Major search engines like Google, Yahoo, and Bing raced to develop harvesting tools and search algorithms that can better leverage structured data. Researchers at Harvard, Cornell, Freie Universität Berlin, and the University of Southampton continue to develop and refine semantic community building and publishing tools.

This paper examines the state of Linked Data tools and infrastructure to determine whether semantic technologies are sufficiently mature for non-expert use, and to identify some of the obstacles to global Linked Data implementation.

Are Semantic Web technologies ready for production?

As more developers have become involved in Linked Data projects (also referred to as Web 3.0 or the

Web of Data) it is perhaps unsurprising that competing ideas have arisen about whether W3C standards like RDF are a defining and necessary component of "Linked Data", or whether a broader definition could be inclusive without diluting the term so much that it becomes meaningless (Berners-Lee, 2006; Cyganiak, 2009; Miller, 2009; Wilde, 2009). Is Linked Data the same thing as the Semantic Web? Do these concepts refer to a specific set of standards? A specific technology stack?

For the purposes of this paper we use the term "Semantic Web" to refer to a full suite of W3C standards including RDF, SPARQL query language, and OWL Web ontology language (W3C, 2010). As for "Linked Data" we will accept the two part definition offered by the research team at Freie Universität Berlin, "The Web of Data is built upon two simple ideas: First, to employ the RDF data model to publish structured data on the Web. Second, to [use http URIs] to set explicit RDF links between data items within different data sources." (Isele, *et al.*, 2009) To determine whether linked data technologies are sufficiently mature for prime time we can explore development and deployment in each of these two separate areas: exposing data as RDF, and linking RDF entities together.

Exposing data as RDF

The necessary first step to enable semantic technologies is for organizations to expose their data using the Resource Description Framework (RDF). Practically this means identifying all of the people, places, things, and concepts that are contained in unstructured text documents, and assigning each of them a unique URI. Each URI must resolve to a document that describes the resource identified by the URI. [Figure 1](#) provides an example of an RDF/XML snippet that is describing an imaginary book. The URI for the book is <http://www.example.com/books/LinkedDataHits>. The other elements are taken from the Dublin Core ontology to describe properties of the book like the author and date of publication.

Figure 1: An example of an RDF/XML snippet describing an imaginary book. This example combines definitions from the W3C RDF namespace, the Dublin Core Elements namespace, and the Dublin Core Terms namespace.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns:dcterms="http://purl.org/dc/terms/">
  <rdf:Description
    rdf:about="http://www.example.com/books/LinkedDataHits">
    <dc:title>Linked Data Greatest Hits</dc:title>
    <dc:creator>Lisa Goddard</dc:creator>
    <dc:contributor>Gillian Byrne</dc:contributor>
    <dc:type>Text</dc:type>
    <dc:publisher>Memorial University</dc:publisher>
    <dcterms:issued>2010</dcterms:issued>
  </rdf:Description>
</rdf:RDF>
```

This kind of highly structured, standardized data will allow search engines to recognize entities (*e.g.*, Microsoft is a company. Bill Gates is a person), to disambiguate concepts (*e.g.*, Windows the software vs. windows the architectural feature), and to perform more precise searches than can be undertaken with current string-matching technology (*e.g.*, Microsoft as the author of a document vs. Microsoft as the subject of a document). When data is structured in a standard way we will be able to search the entire distributed Web with the same power and accuracy that we currently find only in database search engines.

RDF conversion has become a well-defined problem space within the Linked Data community. The process of converting all existing data to RDF can be a major hurdle for organizations with large

numbers of unstructured text documents and few metadata experts. Many tools have been developed to help automate *named entity recognition*, which is the process of using software to automatically identify and classify text elements like the names of persons, organizations, geographical locations, expressions of time, or expressions of quantity. There are several Web services that will organize and encode unstructured text. The Calais Web Service allows a user to paste in a block of unstructured text and returns the major entities, topics, and relationships, which can also be output as RDF. Zemanta offers a term extraction API that provides keyword extraction, auto tagging, and extraction and disambiguation of entities and concepts. BBC Semantic Web architects described how their Muddy Boots system approaches the problem of disambiguation by using similarity clustering based on the context of concepts, "While the term 'apple' is in itself ambiguous, given the context of the terms 'Microsoft' and 'Google', the meaning of 'apple' referring to Apple Inc. becomes clear." [2] The EVRI news indexer and aggregator has released a "sentiment" Web API which extracts entities and then scours the context of each to identify positive and negative sentiments that are associated with that person, place, or thing.

Beyond named entity extraction for unstructured textual data there exist many tools to convert structured data stores to RDF. RDF crosswalks are available to convert common formats into RDF, including e-mail, iCal, spreadsheets, jpegs, and many others (SIMILE, 2008). The D2R Server enables entire relational databases to be exposed as RDF, and enables SPARQL querying against the database content (Bizer and Cyganiak, 2009).

While new tools are constantly being developed to help with the problem of retrospective conversion, there has also been a recent wave of new RDF-compliant publishing tools. Bloggers can incorporate RDF content by using semantic tagging services like those available from Open Calais or Zemanta. Kingsley Idehen (2008) offers a completely semantic blogging solution by installing the popular Wordpress software on top of the Virtuoso platform. Ontowiki and Semantic Media Wiki incorporate Linked Data into collaborative Web authoring software. A number of plug-in modules have been developed to enable automated RDF mark-up for sites using the popular Drupal content management system. The next release, Drupal 7, will incorporate RDF management from the start, and all Drupal 7 pages will be annotated via RDFa (Corlosquet and Clark, 2010).

We are currently witnessing a proliferation of tools to automatically convert existing structured and unstructured text documents to RDF, and the emergence of many new publishing tools that will allow for easy annotation of future documents. Although RDF is a very complex standard, end-user tools and middleware will continue to reduce barriers to RDF implementation in the same way that Web 2.0 interfaces have reduced barriers to participation in online knowledge generation.



Linking RDF entities together

One of the most compelling aspects of the Semantic Web vision is the idea that computers will be able to create new knowledge from existing information. By linking our data to shared ontologies that describe the properties and relationships of objects, we begin to allow machines not just to "understand" content, but also to derive new knowledge by "reasoning" about that content. As a simple example: Linux is an operating system. Operating systems are software. Software is written by people. Therefore Linux is written by people. More than that, a linking hub would allow an intelligent search agent to find a list of Linux developers, see the organizations with which all of those people are affiliated, see the products sold by those organizations, determine which of those products is also software, and produce a list of, for example, all award winning software that has been produced by organizations that employ a Linux developer. In order to enable this kind of sophisticated query, we must develop rich linking hubs to store information about the properties of various entities and the relationships between those entities.

Exposing data as RDF is an important first step, but to actually achieve the linked-data vision we must set explicit RDF links between data items within different data sources. This provides the means by which we can discover more information about a given entity. For example, after the *New York Times* index terms had been converted to RDF, developers explained that "even though we can show you every article written about 'Colbert, Stephen' our databases can't tell you that he was born on May 13, 1964, or that he lost the 2008 Grammy for best spoken word album to Al Gore. To do this we would need to map our subject headings onto other Web databases such as Freebase and DBpedia." (Sandhaus and Larson, 2009)

DBpedia extracts structured information from Wikipedia and makes that data available as RDF. It is one of the largest multi-domain ontologies that currently exist. The knowledgebase automatically evolves as Wikipedia information is updated, and terms mined from Wikipedia have the advantage of representing real community agreement. DBpedia currently contains RDF descriptions of over 2.9 million things, and has emerged as a source of controlled vocabulary for new projects, and a major Semantic Web linking hub.

A 2009 JISC report on learning and teaching found that many of the semantic applications they surveyed either link into, or harvest data from, RDF repositories like DBpedia and Freebase (Tiropanis, 2009). Paul Miller argues that, Web-scale data services such as DBpedia, Freebase, Open Calais and others have much to offer in terms of solutions to constructing and scaling core pieces of data infrastructure. These services have also established a strong lead in assigning and maintaining persistent Web URIs that the community might usefully seek to reuse, instead of inventing new ones [3]. Continued collaborative development of these linking hubs is critical in order to achieve the scale necessary to achieve a high rate of matches across many domains. For example, one BBC attempt to map extracted terms to DBpedia resulted in a success rate of only about 20–30 percent. Developers discovered that many concepts don't have their own Wikipedia article: "TV episodes are often merged into one single list-article, which would be inappropriate to use as an URI for every episode, and many people don't exist in Wikipedia due to low notability." [4] As the BBC develops its own controlled vocabulary it is sharing that data back to DBpedia, and other major semantic service providers like Calais and Zemanta have announced plans to link their entity URIs to DBpedia URIs.

At least one semantic Drupal usability study found that "while linking to external vocabularies was subjectively experienced as easy by all users, a significant time was actually spent deciding to which properties and classes to link with the CCK fields", and the researchers identified a need to "better assist non-Semantic-Web-savvy users in finding the 'right' classes and properties for their needs". [5] The Silk framework is one tool that promises to help discover relationships between data items within different Linked Data sources. It will allow developers to specify which types of RDF links should be discovered between data sources as well as which conditions data items must fulfill in order to be interlinked (Isele, *et al.*, 2010).



Non-technical barriers to broad implementation of Web 3.0

We have established that a technological framework is already in place to support Linked Data production, and that many tools are now available to enable RDF publication and linking by users who are not programmers or metadata specialists. W3C Semantic Web standards have been mature for several years, and real world tools are available for publishing Linked Data; however, only a very small proportion of organizations have made efforts to adopt semantic technologies. Even Tim Berners-Lee admits that the machine-readable Web is still a ways off (Jackson, 2009).

One of the most common criticisms of the Semantic Web vision is that standards like RDF and OWL are difficult to understand conceptually and extremely complex to implement. Part of the Web's success comes from the relative ease of grass roots entry into HTML publishing. RDF, on the other hand, cannot easily be implemented in a lightweight kind of way. Large, well-funded organizations like the BBC and *NYT* may be able to work with the Semantic Web stack from the ground up, but for many the need for investment in a new technology framework (*e.g.*, RDF triple store, SPARQL endpoint) will be a barrier (Miller, *et al.*, 2009a).

Reluctance to adopt new software platforms and productivity tools may be one aspect of the problem, but the lack of people with the right knowledge and skills to implement and use them effectively is at least as much of an obstacle. Few IT professionals are trained in RDF, and not many have experience managing triple stores, or writing SPARQL queries (Jackson, 2009). Even more problematic is a lack of people who understand how to use the tools effectively after implementation. A Norwegian study found that even library staff,

"perceived the magnitude of the organisational challenges facing the NL in implementing the Semantic Web as overwhelming. The basis for this seems to be the perception of the impact of ontology production and maintenance, and the strong focus on semantic meta-data production, all of which have to be established." [6]

If trained cataloguing librarians and metadata experts feel overwhelmed at the prospect of producing semantic metadata and maintaining ontologies, then surely other organizations will find those tasks even more intimidating.

In order to make a large investment of time and money in semantic technologies, organizations have to be convinced that there are costly problems associated with their current suite of technologies, and that semantic technologies will solve these problems and provide a good return on their investment. It is difficult to sell the concept of a single format for Web-based data, for example, when plenty of formats such as relational databases and spreadsheets already annotate data in ways that make it reusable by other systems. Like the idea of a World Wide Web itself was when that was introduced, Ronald Reck posits that "the idea of Linked Data solves a problem we didn't know we had" (Jackson, 2009). In one attempt to quantify the value of information, Michael Bergman estimated that the information contained within U.S. documents represents about a third of total gross domestic product,

or an amount of about US\$3.3 trillion annually, and that the total benefit from improved document access and use to the U.S. economy is on the order of US\$800 billion annually, or about eight percent of GDP (Bergman, 2005). In order to convince decision makers to become part of the Linked Data Web, we will need widespread recognition of the problems associated with current Web and desktop technologies.


Conversely, many people still don't really understand what the Web of Data could accomplish for us. Many Semantic Web evangelists certainly exist including Web architect Sir Tim Berners-Lee; blogger and podcaster Paul Miller from Talis; Nova Spivak, prolific blogger and founder of Radar networks; and, Kingsley Idehen from Virtuoso. For a long time, however, the Semantic Web vision was very much a conversation between specialized technology professionals. The year 2009 marks the first time that the Linked Data conversation has really bubbled up into mainstream media, thanks in part to the U.K. government's recent push to embrace Linked Data. Ultimately it is not enough for technologists to embrace Linked Data — subject specialists, funding agencies, and corporate decision makers must be able to articulate the specific benefits that can be realized within their own expert domains.

Convincing people to invest in the Semantic Web vision is one challenge, but this is by no means the only obstacle to realizing the Web of Data. Even people with deep convictions about Linked Data confess that the systemic obstacles are daunting. We will require a great deal of cooperation and collaboration across institutional and national boundaries if Linked Data is ever to really achieve its potential. We begin to understand the scope of this problem when we consider the difficulty of achieving consensus and consistency across the government departments of a single nation, or even across the departments of a single university! The issue becomes even more complex when we imagine trying to link corporate data. Business intelligence experts understand corporate data as an asset, and are understandably reluctant to share information that is perceived as providing a competitive advantage. Although some major corporations are already experimenting with semantic technologies, most are not currently sharing their RDF data back to linking hubs like DBpedia. Until for-profit companies perceive that there is a net benefit to publicly sharing their data, it is unlikely that this information will be unlocked for the greater good.

Finally, the Semantic Web will be plagued with many of the same problems that we have on the current Web: concerns about privacy and the release of personal information, battles over intellectual property and rights management, and issues surrounding authority. As Paul Miller put it in one podcast, "the Semantic Web will expose all of the problems of the Web like trust, provenance, and reliability (problems which are already very much with us) in a large distributed space" (Miller, *et al.*, 2009b).



Conclusion

There is no question that the obstacles are daunting, but the dazzling promise of the Semantic Web is a compelling reason to continue the work that has already been started. We can participate by creating Web documents using RDF authoring tools, and by using RDF converters to output existing structured data as RDF. These tools are already available, and many of them are suitable for non-experts. Linking hubs and ontologies also exist, but these require a sustained collaborative development effort in order to enrich existing ontologies, and to model new knowledge domains. More than anything else though, we need successful projects to convince people and organizations that the machine-readable Web merits the huge investment of time and money that will be necessary to achieve Tim Berners-Lee's heady vision. 

About the authors

Lisa Goddard is the Acting Associate University Librarian for Information Technology at Memorial University of Newfoundland Libraries.
E-mail: lgoddard [at] mun [dot] ca

Gillian Byrne is the Head of Acquisitions and Serials at Memorial University of Newfoundland Libraries.
E-mail: gbyrne [at] mun [dot] ca

Notes

1. H.M. Government, 2009, p. 28.

2. Kobilarov, *et al.*, 2009, p. 729.
3. Miller, 2010, p. 38.
4. Kobilarov, *et al.*, 2009, p. 730.
5. Corlosquet, *et al.*, 2009, p. 13.
6. Bygstad, *et al.*, 2009, p. 979.

References

Michael K. Bergman, 2005. "Untapped assets: The \$3 trillion value of U.S. enterprise documents," BrightPlanet Corporation white paper (July), 42 pp., and at <http://www.mkbergman.com/82/untapped-assets-the-3-trillion-value-of-us-enterprise-documents>, accessed 19 August 2010.

Tim Berners-Lee, 2006. "The four rules of Linked Data," at <http://www.w3.org/DesignIssues/LinkedData.html>, accessed 19 August 2010.

Chris Bizer and Richard Cyganiak, 2009. "D2R server: Publishing relational databases on the Semantic Web," at <http://www4.wiwiw.fu-berlin.de/bizer/d2r-server/>, accessed 19 August 2010.

Bendik Bygstad, Georghita Ghinea, and Klaeboe Geir-Tore, 2009. "Organisational challenges of the Semantic Web in digital libraries: A Norwegian case study," *Online Information Review*, volume 33, number 5, pp. 973–985. <http://dx.doi.org/10.1108/14684520911001945>

Stephane Corlosquet and Lin Clark, 2010. "The story of RDF in Drupal 7 and what it means for the Web at large," *DrupalCon 2010* (San Francisco), at <http://sf2010.drupal.org/conference/sessions/story-rdf-drupal7-and-what-it-means-web-large>, accessed 19 August 2010.

Stephane Corlosquet, Renaud Delbru, Tim Clark, Axel Polleres, and Stefan Decker, 2009. "Produce and consume Linked Data with Drupal," *Eighth International Semantic Web Conference* (ISWC 2009; Washington, D.C.), at <http://openspring.net/blog/2009/10/22/produce-and-consume-linked-data-with-drupal>, accessed 19 August 2010.

Richard Cyganiak, 2009. "What's in a name? And the Linked Data police," at <http://dowhatimean.net/2009/11/whats-in-a-name-and-the-linked-data-police>, accessed 19 August 2010.

H.M. Government, 2009. *Putting the frontline first: Smarter government* (data.gov.uk). Norwich, U.K.: Stationary Office, and at <http://www.hmg.gov.uk/media/52788/smarter-government-final.pdf>, accessed 19 August 2010.

Kingsley Uyi Idehen, 2008. "Adding Wordpress blogs into the Linked Data Web using Virtuoso," *Kingsley Idehen's Blog Data Space* (10 April), at <http://www.openlinksw.com/blog/~kidehen/?id=1333>, accessed 19 August 2010.

Robert Isele, Anja Jentzsch, Chris Bizer, and Julius Volz, 2010. "Silk — A link discovery framework for the Web of Data," at <http://www4.wiwiw.fu-berlin.de/bizer/silk/>, accessed 19 August 2010.

Joab Jackson, 2009. "Tim Berners-Lee: Machine-readable Web still a ways off," *Government Computer News* (30 October), at <http://gcn.com/articles/2009/10/30/berners-lee-semantic-web.aspx>, accessed 19 August 2010.

Georgi Kobilarov, Tom Scott, Yves Raimond, Silver Oliver, Chris Sizemore, Michael Smethurst, Christian Bizer and Robert Lee, 2009. "Media meets Semantic Web — How the BBC uses DBpedia and Linked Data to make connections," *European Semantic Web Conference — ESWC 2009* (Heraklion, Greece), at <http://www.eswc2009.org/>, accessed 19 August 2010.

Paul Miller, 2010. *Linked Data horizon scan*. Bristol, U.K.: Joint Information Systems Committee (JISC), at <http://linkeddata.jiscpress.org/>, accessed 19 August 2010.

Paul Miller, 2009. "Does Linked Data need RDF?" *Cloud of Data* (19 July), at <http://cloudofdata.com/2009/07/does-linked-data-need-rdf/>, accessed 19 August 2010.

Paul Miller, 2008. "Sir Tim Berners-Lee: Semantic Web is open for business," *ZDNet.com* (26 February), at <http://blogs.zdnet.com/semantic-web/>, accessed 19 August 2010.

Paul Miller, Benjamin Nowack, Greg Boutin, and Tom Tague, 2009a. "October 2009: The Semantic

Web Gang discuss RDF," at <http://semanticgang.talis.com/2009/10/19/october-2009-the-semantic-web-gang-discuss-rdf>, accessed 19 August 2010.

Paul Miller, William Mougayar, Greg Boutin, and Leigh Dodds, 2009b. "November 2009: The Semantic Web Gang discuss Egentia, DBpedia Live, and more," at <http://semanticgang.talis.com/2009/11/20/november-2009-the-semantic-web-gang-discuss-egentia-dbpedialive-and-more/>, accessed 19 August 2010.

Evan Sandhaus and Robert Larson, 2009. "First 5,000 tags released to the Linked Data cloud," *New York Times* (29 October), at <http://open.blogs.nytimes.com/2009/10/29/first-5000-tags-released-to-the-linked-data-cloud/>, accessed August 19 2010.

Semantic Interoperability of Metadata and Information in unLike Environments (SIMILE) Project. MIT, 2008. "RDFizers — SIMILE," Cambridge Mass.: MIT, at <http://simile.mit.edu/wiki/RDFizers>, accessed 19 August 2010.

Thanassis Tiropanis, Hugh Davis, David Millard, Mark Weal, Su White, and Gary Wills, 2009. *Semantic technologies in learning & teaching*. Bristol, U.K.: Joint Information Steering Committee (JISC), at <http://www.jisc.ac.uk/publications/reports/2009/semantictechnologiesreport.aspx>, accessed 19 August, 2010

Erik Wilde, 2009. "The Linked Data™ police," *Dretblog* (20 November), at <http://dret.typepad.com/dretblog/2009/11/the-linked-data-police.html>, accessed 19 August 2010.

World Wide Web Consortium (W3C), 2010. "Semantic Web," at <http://www.w3.org/standards/semanticweb/>, accessed 8 October 2010.

Editorial history

Received 19 August 2010; accepted 22 October 2010.



"Linked Data tools: Semantic Web for the masses" by Lisa Goddard and Gillian Byrne is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License](http://creativecommons.org/licenses/by-nc-nd/3.0/).

Linked Data tools: Semantic Web for the masses
by Lisa Goddard and Gillian Byrne.

First Monday, Volume 15, Number 11 - 1 November 2010

<http://firstmonday.org/ojs/index.php/fm/rt/prINTERfriendly/3120/2633>