The Semantic Web: An introduction for information professionals

Matt Moore[*]

The "Semantic Web" is a term that has been in common use for a decade. This article examines what the Semantic Web means for information professionals and provides an overview of some of the core technologies, such as RDF. It then explores the network of linked data that has arisen from using these technologies, before concluding with three suggestions for information professionals wanting to explore and exploit the Semantic Web for their own work.

What is the Semantic Web?

> Most of the Web's content today is designed for humans to read, not for computer programs to manipulate meaningfully.[1]

The world wide web – the marrying of the infrastructure of the internet and the interactivity of hypertext – is now well established. The indexed web contains over 13 billion pages.[2] However, for many computer scientists the web is only a partial success. The information contained on these 13 billion plus pages is fragmented, isolated and ambiguous. The Wikipedia page for "Sydney" provides a wealth of information on the State capital of New South Wales but do we know if this is the same "Sydney" as is serviced by the QANTAS and Jazz airline services? The answer in this case is "yes" and "no" respectively. Jazz is a Canadian airline and the "Sydney" it services is in Nova Scotia. Someone browsing the web would have to assemble these different "Sydney" resources themselves and decide what could be correctly placed together and what should be kept apart.

Some imagine a different kind of world wide web. The use of "Sydney" on a page would be clearly marked as to whether it referred to a town in Nova Scotia or a region in Australia – and whether the Australian "Sydney" meant the local government "City of Sydney" or the greater metropolitan area of "Sydney" or some other variation. This detailed, precise metadata would allow machines to browse the web "intelligently" and interact with the content available. Tim Berners-Lee was instrumental in the development of the world wide web and he has been the most vocal proponent of a semantic successor. Here he describes a fictional example of this "Semantic Web" in a 2001 article:

> At the doctor's office, Lucy instructed her Semantic Web agent through her handheld Web browser. The agent promptly retrieved information about Mom's prescribed treatment from the doctor's agent, looked up several lists of providers, and checked for the ones in-plan for Mom's insurance within a 20-mile radius of her home and with a rating of excellent or very good on trusted rating services. It then began trying to find a match between available appointment times (supplied by the agents of individual providers through their Web sites) and Pete's and Lucy's busy schedules.[3]

In this world, information has to be made available in a more sophisticated form than flat text and the tools to parse it have to be equally sophisticated. Above all, this Semantic Web requires international standards. Much work in this domain has been carried out by the World Wide Web

[*] Matt Moore is a director of Innotecture, an occasional lecturer at the University of Technology, Sydney and chair of the New South Wales Knowledge Management Forum. He has spent over a decade working in knowledge and information management, learning and development and internal communications with organisations such as PwC, IBM, Oracle and the Australian Securities and Investment Commission.

[1] Berners-Lee T, Hendler J and Lassila O, "The Semantic Web", *Scientific American Magazine* (17 May 2001), http://www.scientificamerican.com/article.cfm?id=the-semantic-web viewed 15 March 2011.

[2] De Kunder M, "The Size of the World Wide Web", *WorldWideWebSize* (14 March 2011), http://www.worldwidewebsize.com viewed 15 March 2011.

[3] Berners-Lee et al, n 1.

Consortium (W3C).[4] Further below, this article will outline some of the critical technologies that have been developed by the W3C (as well as some that have not).

Whether the Semantic Web has arrived yet, or will arrive in the future, is a matter of some debate. A recent report by Pew Research Center[5] asked nearly 900 internet technologists, academics and opinion leaders for their thoughts on the Semantic Web. The results were diverse – 47% of them agreed with the following statement: "By 2020, the semantic web envisioned by Tim Berners-Lee will not be as fully effective as its creators hoped and average users will not have noticed much of a difference."

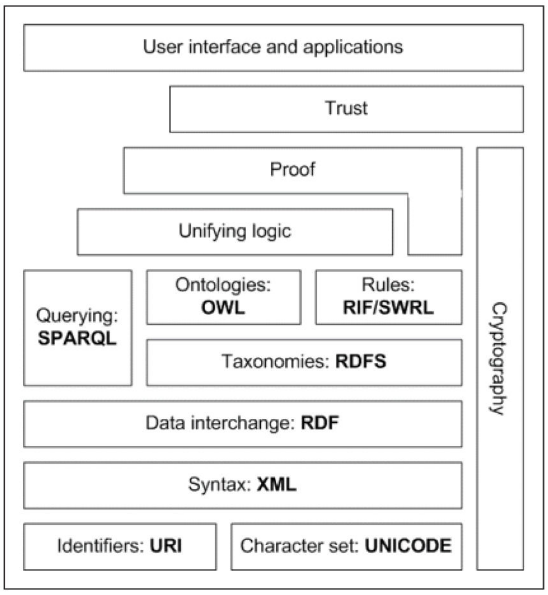
In some ways the efforts of those who advocate the Semantic Web mirror those who have argued for a global political or economic structure with international law and integrated markets. All of these are big jobs that will take decades (if they succeed at all). For each of them, there is plenty of visionary rhetoric that plays out as a messy reality. There are those who believe that there needs to be a single, unified approach to integration, whereas others prefer smaller, piecemeal projects; some are simply sceptical of the whole enterprise.

What is clear is that the internet is now more "semantic" than it was a decade ago. Some of these new technologies will now be discussed.

THE SEMANTIC STACK

One way of understanding the technologies that make up the Semantic Web is as a set of layers. At the bottom are foundational technologies and each layer builds upon the previous one. This is something of a simplification but let's start with the simple.

FIGURE 1 The semantic stack



Source: http://www.semanticweb.org.

[4] Herman I, "The State of the Semantic Web" (Speech delivered at the Semantic Technology Conference, San Jose, 18 May 2008), http://www.w3.org/2008/Talks/0518-SanJose-IH viewed 15 March 2011; Herman I, "How Does the Semantic Web Work?" (Speech delivered at the Semantic Café event, Sao Paulo, Brazil, 15 October 2010), http://www.w3.org/2010/Talks/1015-SaoPaulo-SemCafe-IH/#talk viewed 15 March 2011.

[5] Anderson J and Rainie L, "The Fate of the Semantic Web", *Pew Research Center* (4 May 2010), http://www.pewinternet.org/Reports/2010/Semantic-Web.aspx viewed 15 March 2011.

One of the two foundational technologies is the character set. The character set encodes textual elements as a string of data. There are different character-set standards. The Unicode standard attempts to provide a standard that is compatible across different character sets (eg the Roman and Cyrillic alphabets). For example, the UTF-8 version of Unicode represents the symbol "A" in the Roman alphabet as hexadecimal code 0041. Without character representation standards, there would be no text on the world wide web nor would it be easy to transfer files between text creating programs such as Microsoft Word and Google Docs.

The identifier points towards a specific resource. The URI (Uniform Resource Identifier) standard includes both URLs (Uniform Resource Locators) and URNs (Uniform Resource Names). An example of a URL is http://www.en.wikipedia.org/wiki/Uniform_Resource_Locator ("http:" being the scheme; "en.wikipedia.org" being the domain; and "wiki/Uniform_Resource_Locator" being the path). An example of a URN would be "urn:issn:0167-6423", which refers to the *Science of Computer Programming* journal, identified by its ISSN number.

The commonly used metaphor to compare a URN to a URL would be the name of a person and his or her address respectively (although in most countries, a name is not sufficiently unique when dealing with individuals – so it is often paired with a date of birth or replaced with a driver's licence or social security number). Identifier standards build on pre-existing structures, eg the ISSN (International Standard Serial Number) system, and as such they should be a familiar concept to information professionals. Identifier standards are critical and without them the world wide web could not exist in a usable system of links.

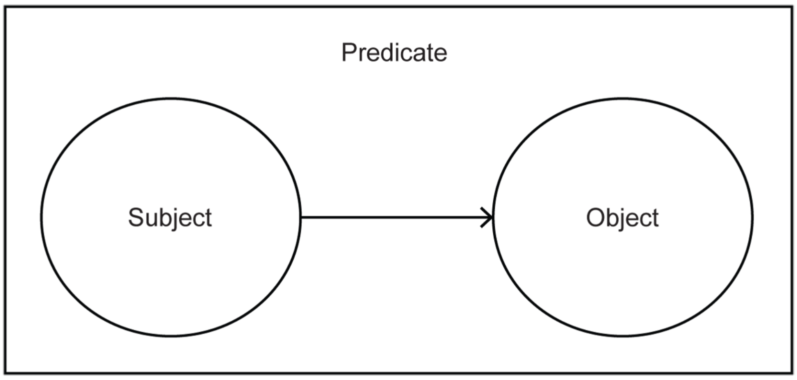
The syntax represents structured information such as documents, transactions etc. XML (eXtensible Markup Language) is a simple, text-based format for syntax representation. At first glance, XML looks like the HTML (HyperText Markup Language) that makes up webpages. Both XML and HTML documents consists of content and markup, eg <a href="http://.en.wikipedia.org/wiki/XML">Wikipedia page on XML</a> consists of the phrase "Wikipedia page on XML" marked up with a tag (anything surrounded by < and > is a tag) that links this phrase to a relevant page on Wikipedia. This involves text (in Unicode) and an identifier (the URL).

However, there are important differences between HTML and XML. HTML is a language used to present information on webpages. XML is a framework for creating applications that can manage information in a structured way. The markup elements in an HTML document are fixed, whereas the elements in an XML document can be customised depending on the situation. You can have tags that relate to <customer>, <process> or <location>. XML applications include: RSS (Really Simple Syndication), used to provide update feeds for websites and blogs; KML (Keyhole Markup Language), used to annotate locations in Google Earth; and RDF (Resource Description Framework), described below. Most casual users of the internet are blissfully unaware of the ubiquity of XML.

RDF (Resource Descriptor Framework) was originally developed as a means to describe webpages but has evolved into a framework that can represent more generalised relationships between entities.[6] These relationships are expressed as "triples" containing a subject, predicate and object. Triples can be represented in a graph format with the subjects and objects as nodes and the predicate as a linking arc.

[6] *RDF Primer* (World Wide Web Consortium, 2004), http://www.w3.org/TR/rdf-primer viewed 15 March 2011.

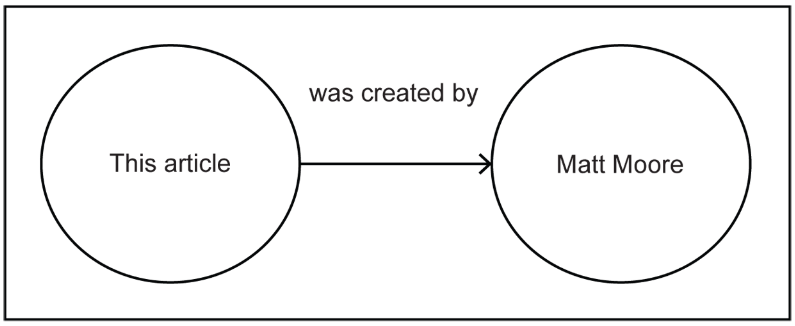
FIGURE 2 The RDF triple structure



An example of the structure of a triple would be: "This article was created by Matt Moore". In this example:

- "This article" is the subject;
- "was created by" is the predicate; and
- "Matt Moore" is the object.

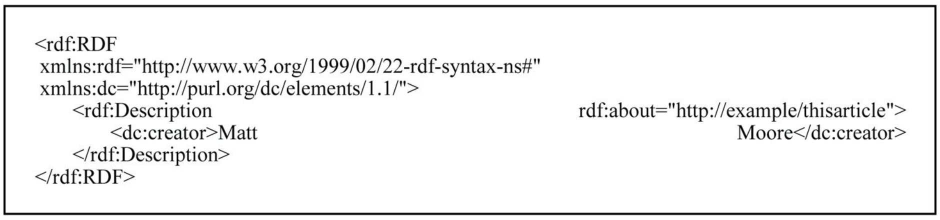
FIGURE 3 Example of an RDF triple



No RDF triple statement stands by itself. The predicate must always be a URI that refers to an existing resource (subjects and objects may also be URIs). For example, in the above statement "was created by" could be represented by the following URI: http://www.purl.org/dc/elements/1.1/creator, which refers to the Creator element within the Dublin Core Metadata Element Set.

Triples can be written using different forms of syntax. As noted above, one such syntax is XML. However, RDF statements written in XML can be lengthy so more compact forms such as Notation3 (N3) can be used. Figures 4 and 5 show XML and N3 versions of the same statements (based on the previous example).

FIGURE 4 Example of an RDF triple using RDF/XML



```
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dc="http://purl.org/dc/elements/1.1/">
      <rdf:Description                                        rdf:about="http://example/thisarticle">
           <dc:creator>Matt                                                       Moore</dc:creator>
      </rdf:Description>
</rdf:RDF>
```

FIGURE 5 Example of an RDF triple using N3

```
<http://example/thisarticle> <http://purl.org/dc/elements/1.1/title> "Matt Moore"
```

RDFS (RDF Schema) expands on RDF by introducing the concepts of classes and properties. "Matt Moore" in the previous example might belong to the foaf:Person class. "foaf" stands for Friend Of A Friend, a vocabulary used to describe people and their relationships to each other. Properties describe relationships between subject and object resources – so that vendors might be specified as being companies.

OWL (Web Ontology Language) adds even more vocabulary for describing properties and classes: among others, relations between classes (eg disjointness); cardinality (eg "exactly one"); equality; richer typing of properties; characteristics of properties (eg symmetry); and enumerated classes. OWL can be used to create sophisticated relationships between entities.

SPARQL (SPARQL Protocol and RDF Query Language) is a query language that allows the search and retrieval of information and relationships held in an ontology. It could be used to find out what else this entity called "Matt Moore" is responsible for and those other entities that have relationships with it.

There are a range of technologies that exist outside and alongside the W3C semantic stack.

"Designed for humans first and machines second", microformats have evolved as tactical solutions to information structure problems on the web rather than as attempts to create new standards. For example, hCard is a simple format for representing people, places and organisations using an adapted form of HTML. Research by Google in 2010 indicates that microformats such as hCard are an order of magnitude more popular than RDF-based equivalents[7] – however, over 95% of sites surveyed contained neither.

Topic Maps are another approach to describing information resources and relationships through graphs. These have been documented as ISO 13250. Similar to mind maps and concepts maps, they are primarily for use by humans rather than machines.

Linked Data

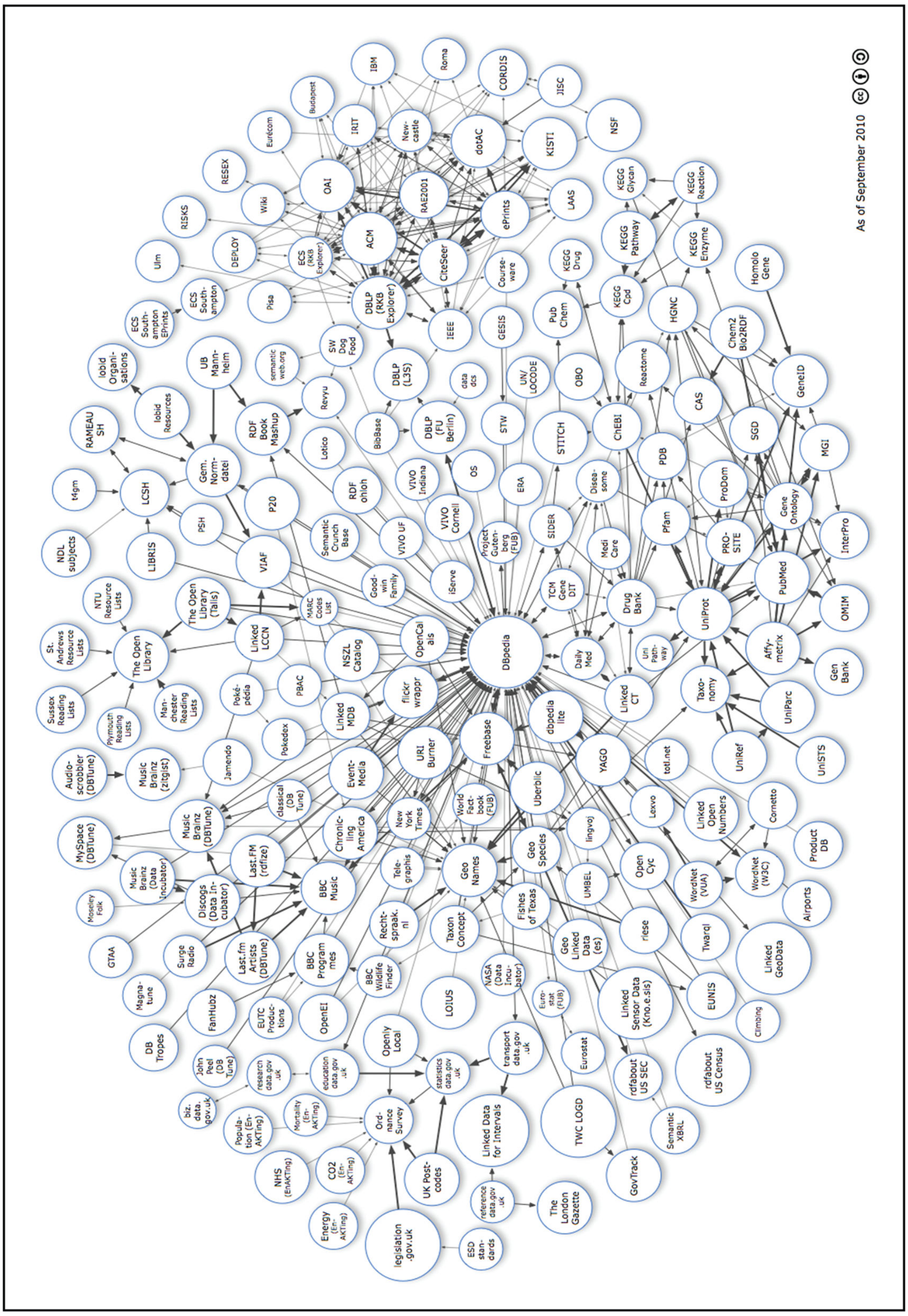
The proliferation of overlapping and competing semantic technologies for the internet is of interest to computer scientists and web developers but why should the rest of us care? Tim Berners-Lee had a vision of intelligent agents interacting with smart web data but we still do not experience that so what do we have? At the moment, what we have is "Linked Data".

The Linking Open Data (http://www.linkeddata.org) project began in 2007 with the aim of exposing and linking datasets of RDF triples on the internet using standards such as RDF and SPARQL.[8] The early datasets were from small organisations and universities but larger organisations, such as the BBC, Thomson Reuters and the United States government, have now created their own. The map below provides an overview of the Linked Data cloud.

[7] MacManus R, "Google's Semantic Web Push: Rich Snippets Usage Growing", *ReadWriteWeb* (24 June 2010), http://www.readwriteweb.com/archives/google_semantic_web_push_rich_snippets_usage_grow.php viewed 15 March 2011.

[8] Bizer C, Heath T and Berners-Lee T, "Linked Data – The Story So Far" (2009) *International Journal on Semantic Web and Information Systems*, http://www.tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf viewed 15 March 2011.

FIGURE 6 Linking Open Data cloud diagram



Source: Cyganiak R and Jentzsch A, http://richard.cyganiak.de/2007/10/lod/.

The scale of Linked Data is impressive:

- The United States government's http://www.data.gov offers over four billion RDF triples.
- DBpedia is based on information held within Wikpedia. It covers 12,000 persons, 413,000 places and 146,000 species. It now has over a billion triples.
- Linked GeoData provides the spatial data from OpenStreetMap. It has over two billion triples.

The "linked" part of Linked Data is as important as the "data" aspect. Many entities appear in multiple datasets. Sometimes the same URI is used across datasets (eg ISBN numbers in publishing). Sometimes different datasets use different URIs – eg DBpedia uses the URI http://www.dbpedia.org/resource/Berlin to identify Berlin, while Geonames uses the URI http://www.sws.geonames.org/2950159 to identify Berlin. If no shared naming schema for entities exist then RDF links can be generated algorithmically based on the similarity of entities within both data sets. For example, RDF links between artists in two music data sets were created with a similarity metric that compared the names of artists as well as the titles of their albums and songs.

So what uses are these resources being put to? Two examples from the United Kingdom indicate what can be done with Linked Data.

Project LUCERNO at the Open University

The Open University (OU) is a British distance learning and research institution with over 150,000 students. Currently, a student wishing to discover all of the material – books, DVDs, CDs, TV programs, Podcasts, Open Educational Resources etc – related to a specific OU course (aka module), would have to consult a different data source, with a different system and interface, for each type of resource required, explore their results and integrate them manually. In a similar scenario, the same resources are needed by lecturers in creating new courses or tutorials, as well as by researchers in connecting the result of their research to existing resources.

The LUCERO (Linking University Content for Education and Research Online) Project at the OU is investigating and prototyping the use of Linked Data technologies and approaches to linking and exposing data for students and researchers.

LUCERO is working on exposing a number of OU datasets as Linked Data, including: Course Information Research Publications recorded in the Open University Research Online (ORO) repository; people information (specifically OU staff); podcasts; course material metadata – this is descriptive information from the OU library catalogue covering books and other media (eg DVDs, CDs) but excluding online material; and a number of OU research and teaching material resources.

The Open University hopes to enable students to find all the material directly related to their course/module as well as supplementary material based on direct links, cross-repositories such as subject classifications, as well as based on indirect (and possibly external) links such as "the people involved in the creation of the resources".[9] Students would also be able to find relevant material stocked in other, nearby institutions.

Semantic Web at the BBC

The BBC is the largest broadcasting corporation in the world and it publishes extensive amounts of content online, as text, audio and video. Historically, the website has focused on supporting broadcast brands (eg Top Gear) and a series of domain-specific sites (eg news, food, gardening etc). That is, the focus has been on providing separate, standalone HTML sites designed to be accessed with a desktop web browser. These sites can be very successful, but tend not to link together, and so are less useful when people have interests that span program brands or domains.

BBC Programmes was launched in the summer of 2007. Its goal is to provide a web identifier, with associated HTML pages and machine-readable feeds (RDF/XML, JSON and XML), for every program the BBC broadcasts – allowing other teams within the BBC to incorporate those pages into

[9] Stephens O, *Use Case Collecting Material Related to Courses at The Open University* (Open University, Milton Keynes, 2010), http://www.w3.org/2005/Incubator/lld/wiki/index.php?title=Use_Case_Collecting_material_related_to_courses_at_The_Open_University&oldid=2257 viewed 15 March 2011.

new and existing program support sites, TV Channel and Radio Station sites, and cross program genre sites such as food, music and natural history. BBC Music follows the same principles as BBC Programmes, and provides a web identifier for every artist the BBC has an interest in (featured in music programs, in BBC events etc). BBC Music is underpinned by the Musicbrainz music database and Wikipedia, thereby linking out into the web as well as improving links within the BBC site. BBC Music takes the approach that the web itself is its content management system. Site editors directly contribute to Musicbrainz and Wikipedia, and BBC Music will show an aggregated view of this information, put in a BBC context. As another example, the BBC Wildlife Finder provides a web identifier for every species (and other biological ranks), habitat and adaptation the BBC has an interest in. BBC Nature aggregates data from different sources, including Wikipedia, the WWF's Wildfinder, the IUCN's Red List of Threatened Species, the Zoological Society of London's EDGE of Existence program, and the Animal Diversity Web. BBC Wildlife Finder repurposes that data and puts it in a BBC context, linking out to program clips extracted from the BBC's Natural History Unit archive.[10]

Creating web identifiers for every item the BBC has an interest in, and considering those as aggregations of BBC content about that item, allows us to enable rich cross-domain user journeys. This means BBC content can be discovered by users in many different ways, and content teams within the organisation have a focal point around which to organise their content. The approach has also proved to be an efficient one – allowing different development teams to concentrate on different domains while at the same time benefiting from the activities of the other teams. The small pieces loosely joined approach, which is manifest in any Linked Data project, significantly reduces the need to coordinate teams while at the same time allowing each team to benefit from the activities of others.

HOW SERIOUSLY SHOULD INFORMATION PROFESSIONALS TAKE THE SEMANTIC WEB?

The Semantic Web has so far failed to live up to the initial hype surrounding it at the turn of the millennium. However, before we dismiss it, we should remember that most innovations fail to live up to their hype in the short term, and that, for a project as ambitious as this, the "short term" stretches into the coming decade. On the other hand, their potential complexity, means that Semantic Web projects can absorb a large amount of time, effort and money if you so choose. The discerning information professional should consider the following opportunities.

Understand the basics of the various Semantic Web technologies

Spend some time understanding the layout and logic of RDF triples, OWL ontologies and microformats. They are often expressed in forbidding formats but underneath many familiar concepts are present. In some cases, such as the Dublin Core, the principles that are expressed overlap with traditional library and information management work. There are a growing number of free or cheap tools that allow you to build and navigate Semantic Web structures. This basic knowledge and experience will allow you to have more effective conversations with technologists when the time comes.

Find what already exists

The power of linked data is in reuse. A canny early move would be to find what you can reuse to make your life easier. Can an existing dataset enrich the information that you already hold in some form? The Powerhouse Museum in Sydney is using Linked Data via Open Calais to enrich the online details of its exhibits (names of famous individuals are linked to entries on a global biographical database). This opportunity should be of special interest to smaller organisations that may not have the resources to create extensive metadata records.

Put your own material out there

The release of the Government 2.0 Taskforce report has given added impetus to public sector organisations to make their information available, and bodies such as the Australian Bureau of

[10] Raimond Y, Scott T, Sinclair P, Miller L, Betts S and McNamara F, "Case Study: Use of Semantic Web Technologies on the BBC Web Sites", *BBC* (2010), http://www.w3.org/2001/sw/sweo/public/UseCases/BBC viewed 15 March 2011.

Statistics have been leading the way. However, this information needs to be structured in ways that stakeholders can easily consume. Even if you work in a private organisation there is still an incentive to start exposing your schemas. By doing so, you are stating your view of the world and others will probably pick it up for the sake of convenience. This is a powerful position to be in as you are defining the territory on which you operate. Who would want to miss out on such an opportunity?