The semantic web and the digital library

Mary Burke
School of Information and Library Studies, University College Dublin, Dublin, Ireland

Received 27 September 2008
Revised 4 March 2009
Accepted 6 March 2009

Abstract

Purpose – The purpose of this paper is to discuss alternative definitions of and approaches to the semantic web. It aims to clarify the relationship between the semantic web, Web 2.0 and Library 2.0.

Design/methodology/approach – The paper is based on a literature review and evaluation of systems with semantic web features. It identifies and describes semantic web projects of relevance to libraries and evaluates the usefulness of JeromeDL and other social semantic digital library systems. It discusses actual and potential applications for libraries and makes recommendations for actions needed by researchers and practitioners.

Findings – The paper concludes that the library community has a lot to offer to, and benefit from, the semantic web, but there is limited interest in the library community. It recommends that there be greater collaboration between semantic web researchers and project developers, library management systems providers and the library community. Librarians should get involved in the development of semantic web standards, for example, metadata and taxonomies.

Originality/value – The paper clarifies the distinction between semantic web and Web 2.0 in a digital library environment. It evaluates and predicts future developments for operational systems.

Keywords Semantics, Worldwide web, Digital libraries

Paper type General review

Introduction

The semantic web fulfils Tim Berner-Lee's original concept of the web:

> My dream for the Web has two parts. In the first, I see the Web becoming a much more powerful means for collaboration among people. In the second, collaborations extend to computers. Machines become capable of analyzing all the data on the Web – the contents, links and transactions between people and computers (Anbarasan, 2000).

The semantic web converts web pages from being readable and displayable by computers to being understandable by computers. It does this by adding extra metadata to web pages and by sharing this metadata between multiple applications. It enables computers to understand a web page in the way a human does, so that computers can find, share and integrate information on the web. It depends on two pillars: metadata and ontologies. Metadata provide the ability to identify and exploit relationships between items. Ontologies enable equivalences to be created between items in different collections which have been described using different vocabularies.

This paper identifies the components of the semantic web and evaluates the relationship between the semantic web and Web 2.0 and Library 2.0. It identifies and describes semantic web projects of relevance to libraries. It evaluates the usefulness of social semantic library systems, including JeromeDL. It discusses actual and potential applications for libraries, and identifies what researchers and practitioners need to do in order to participate in the semantic web.

Aslib Proceedings: New Information Perspectives
Vol. 61 No. 3, 2009
pp. 316-322

0001-253X
DOI 10.1108/00012530910959844

Semantic web building blocks

Hypertext Markup Language (HTML), the mark-up standard for web pages, is a display format that does not record the underlying structure of the data. The semantic web uses the resource description format (RDF) to describe data models, including objects and relationships between objects. The range of objects is very wide and includes people, meetings and products, in addition to text, image and audio documents. RDF converters convert from application specific formats, for example, MARC, JPEG, XML, EXCEL, Flickr, etc., to RDF (Herman, 2007).

OWL web ontology language adds more vocabulary for describing properties and classes. Ontologies facilitate integration of resources where different identifiers are used for the same concept, for example, "author" and "creator". Extensible Markup Language (XML) provides a basic syntax for content structure. It does not deal with meaning. XML Schema is a language for describing the structure and content of elements in XML documents. SPARQL is a protocol and query language for resources in the semantic web.

Semantic web applications

There are many successful semantic web projects in a variety of sectors including commerce, medicine, and scientific publishing. Some key semantic web projects will now be described.

Friend of a friend (FOAF (www.foaf-project.org/)), is a machine-readable ontology for describing people, their activities and their relations to other people and objects using RDF and OWL. It can be integrated with library systems to incorporate bookmarks and annotations by people identified by a user as "friends".

Semantically interlinked online communities (SIOC (http://sioc-project.org/)) provides a vocabulary of terms and relationships that model data spaces, for example, blogs, mailing lists and image galleries. It enables these resources to be linked and the content to be retrievable.

Linking open data project is a community lead project to connect RDF datasets on the web (Cyganiak, 2007). One of its recent datasets is DBpedia, based on Wikipedia (http://wiki.dbpedia.org/About). Users may search the data and link to other datasets on the web. DBpedia describes almost two million "things" including places, people, music albums, and films. The dataset also has over 2,500,000 links to images, external web pages, and other RDF datasets (Cyganiak, 2007).

Musicbrainz (http//musicbrainz.org) is a large database of music metadata for over half a million albums with active links to other types of music information. It enables users to download and exchange metadata for music collections. It uses RDF and uniform resource identifiers (URIs) to enable links to relevant music web pages, for example, discographies, biographies, etc.

The key feature of these semantic web applications is their ability to link data objects in multiple databases and thereby facilitate data integration. They are effective for specific subject areas because of the quality of the existing metadata and the willingness of the user community to enhance metadata and develop ontologies.

Web 2.0

Many semantic web applications also involve Web 2.0 and social networking. Some writers use the terms semantic web and Web 2.0 interchangeably so this paper will

attempt to clarify the relationship between them. The O'Reilly XML (Zambonini, 2005) and the ZDNet Blogs (MacManus, 2005) provide a good debate on semantic web versus Web 2.0.

Web 2.0 is an informal flexible way of integrating disparate web services. It requires less dependence on shared vocabularies and provides workable rather than totally perfect solutions. It focuses on the user interface rather than on technology or standards. It encourages users of information services to create, share, collaborate and interact. Web 2.0 applications include social-networking sites (e.g., MySpace (www.myspace.com/), Bebo (www.bebo.com/), Facebook (www.facebook.com/)), wikis (e.g., Wikipedia (http://wiki.dbpedia.org/About)), photo-sharing sites (e.g. Picasaweb (http://picasaweb.google.com), Flickr (www.flickr.com)), and bookmarking sites with collaborative tagging (e.g. Connotea[1]), Delicious (http://delicio.us/)).

Shared sites that allow tagging or allow for creation of metadata by multiple users contribute to the semantic web by enriching metadata. They provide a user driven approach to semantic web development, creating the social semantic web.

Web 2.0 may be seen as the direct opposite to the semantic web with the emphasis of the latter on standardisation, creation of perfect metadata, and using computers to link the metadata. There is general acceptance that the two approaches are complementary and can be integrated, with a semantic web back-end and a Web 2.0 interface. Decker (2005) clarifies the distinction:

> ... the Semantic Web effort itself does not provide applications ... it rather provides standards to interlink applications. So for Web 2.0 (or the Semantic Web 2.0) Semantic Web recommendations provide a way to interlink applications.

Another active area of integration of semantic web and Web 2.0 is the use of SPARQL query language, which was developed as a query language for RDF in semantic web, for searching Web 2.0 applications. Clark envisages using SPARQL to arbitrarily slice the data of Flickr, Delicious, Google, and a range of Web 2.0 sites, all FOAF files, and MusicBrainz, etc., and convert these results into an RSS 1.0 feed (Clark, 2005).

The semantic web and Web 2.0 are complimentary, where the semantic web provides for intelligent linking and use of resources which are then accessible to more users through a Web 2.0 interface. The user tagging of items in Web 2.0 can be linked to more formal vocabularies in semantic web. The most fertile area for future development lies in projects that integrate semantic web and Web 2.0, including a number of projects at the Digital Enterprise Research Institute (DERI – www.deri.ie/) to semantically interlink community sites, a semantic blog, and the Social Semantic Desktop (Decker, 2005).

Tom Gruber uses the term "collective knowledge systems" to describe applications which unlock the "collective intelligence" of the social web with knowledge representation and reasoning techniques of the semantic web (Gruber, 2008).

The phrase Library 2.0 was coined in late 2005 by Michael Casey in his blog LibraryCrunch (Casey, 2005). Library 2.0 takes the concept of user evaluation further by applying Web 2.0 features which enable users to share and annotate library resources in electronic teaching and research environments.

An exciting development for the future is the Wikipedia 3.0 proposal which would enrich Wikipedia through the addition of semantic web features. By giving semantic web tools to the Wikipedia community, including librarians, it would be possible for

the user community to help develop domain specific ontologies. Wikipedia would then become a real rival to Google, providing answers to questions instead of links to web pages containing keywords (Fawzi, 2006). Other projects that introduce semantic web features to Wikipedia include DBpedia (http://wiki.dbpedia.org) and Semantic MediaWiki (http://semanticweb.org/wiki/Semantic_MediaWiki).

Semantic web and libraries

Baker (2006) in his review of "Digital library futures: a UK HE and FE perspective" sees personalisation of digital libraries as important. Both semantic web and Web 2.0 facilitate personalisation. Ferran presents a model for how personalisation could be achieved using ontologies (Ferran *et al.*, 2005). These are built using sub-ontologies which describe the basic components of the personalisation system, including users, digital resources, actions, and navigational profiles.

Campbell and Fast (2004) consider the potential of semantic web for future catalogues in academic research libraries. They recognize that interoperable transfer of metadata in metadata harvesting programmes is important, for example, the Open Archives Initiative Protocol for Metadata Harvesting (www.openarchives.org/pmh/). The basic catalogue record can be enhanced with information gathered by a software agent from RDF encoded resources on the web. These might include the work's genre, its historical and intellectual content, information about the author, and bibliographic features. Campbell and Fast assume that cataloguers will spend time locating RDF resources, assessing their reliability and usefulness and participating in mark-up projects to make library catalogue material RDF encoded and available.

JeromeDL (www.jeromedl.org) is a social semantic digital library system which incorporates many features of semantic web and Web 2.0 in an operational system. It was developed by the semantic web research group (http://sw.deri.ie) at the Digital Enterprise Research Institute (DERI). The features of this system include individualisation of user profiles, the ability for users to annotate items, create personal bookshelves, advanced query facilities, etc. JeromeDL describes each resource using three types of metadata: structure, bibliographic and community and delivers services using each of these metadata types. It integrates well with existing library practice by allowing librarians to describe resources using a range of controlled vocabularies that include: authority files, with a list of authors, editors and publishers; classification schemes, for example, Dewey, and WordNet dictionary for keywords.

Structural information about resources, for example, chapters in a book, is recorded using an ontology in RDF. The system has a bibliographic ontology based on Dublin Core data and a structure ontology. The bibliographic ontology uses MarcOnt as a mediation standard between MARC21, Bibtex and Dublin Core. Individual librarians can suggest new concepts for the ontology. A secure snapshot repository with RDF as the common data model allows users to easily integrate information from different sources with appropriate protection of sensitive data, for example, passwords. Semantic Query Expansion is available to refine a query based on statistical analysis. Person concepts are expanded using FOAF. Thesaurus concepts are refined with narrower and broader relationships and keywords are expanded using WordNet pointers like synonyms.

JeromeDL delivers a direct RDF query service as a mash-up for searching other libraries and external services. There is a natural language query template. Users may also filter resources by various criteria including author, keywords, topics and publication

type. FOAFRealm allows users to control their profile information and it also manages an authority list of authors, editors and publishers. Social tagging is encouraged using semantic keywords from the WordNet vocabulary, however a user is also able to tag a resource with keywords that are not in WordNet. Other users can see these tags. JeromeDL has been implemented in DERI in Ireland and in Gdansk University in Poland.

Building Resources for Integrated Cultural Knowledge Services (BRICKS , www.brickscommunity.org) is an EU Sixth Framework (FP6) project for sharing and integrating digital cultural heritage collections using open source software. There are a number of implementations of BRICKS, including the Austrian National Library's Pictures Archive and the Italian Consorzio Forma's image collection.

The Alexandria Digital Library (www.alexandria.ucsb.edu) is a distributed digital library with geo-referenced features. Its semantic web features include the ability to make intelligent use of geographical coordinates.

The Talia digital library platform has many similarities with JeromeDL but focuses more on scholarly research and publishing in the humanities and social sciences (Nucci *et al.*, 2008).

There is significant research activity at European level in developing semantic web applications for digital libraries. The projects listed above have received funding from FP6, FP7 Digital Libraries, e-Content, e-Contentplus, and other EU programmes (http://cordis.europa.eu/fp7/ict/). The European Digital Library has been extended to Europeana (www.europeana.eu/home.php), with the addition of semantic web and Web 2.0 features for user friendly access to European cultural resources.

Conclusions

The semantic web provides exciting potential for extension of library web services but with a high overhead of time and resources for the creation and maintenance of metadata and taxonomies. A key criterion for future success is the ability of the projects to generate metadata and ontologies automatically for scalable solutions. Evaluation and testing of semantic web systems with real users in a variety of different institutions is essential. There appears to be potential to integrate the resource and ontology expertise of subject portals into semantic web digital library solutions. While we are likely to see increasing use of semantic web features, complete semantic web implementation will be slower to achieve, given the loyalty of many libraries to their current library management systems. It is encouraging that the widely used Talis library system provides semantic web capability through Talis Platform (www.talis.com/platform/). This works like a shared database capable of storing and searching both content and RDF metadata.

Librarians have considerable expertise in the key components of the semantic web: cataloguing and classification. Their skills include understanding the complexity of information resources, and experience in developing and implementing metadata standards and controlled vocabulary. They also have strong professional commitment to the protection of freedom of access and user privacy, which are seriously threatened by the interconnectivity of the semantic web. The privacy issues arise as the semantic web will make it easier for central agencies to retrieve, interconnect and understand items. There would also be some loss of anonymity for blog entries, for example, through linking of records in different places. Librarians and system developers must

ensure that the semantic web benefits from the expertise of library professionals in these areas.

Greenberg identifies similarities between libraries and the semantic web and presents strong arguments for the applicability of library functions to the semantic web. "Collection development" translates to "semantic web selection"; "cataloging" to "semantic web 'semantic' representation"; "reference" to "semantic web service"; and "circulation" to "semantic web resource use" (Greenberg, 2007).

The library community, in common with many other communities, has expressed considerably more interest in Web 2.0 than in semantic web applications. A recent summit on the semantic web and digital libraries in Ireland (http://wki.corrib.deri.ie/index.php/SemDL/IrishDLSummit) attracted mainly academic researchers and very few members of the library community. Similar international gatherings attract computer scientists rather than librarians, for example, the first workshop on semantic interoperability in the European Digital Library (http://multimedia.semanticweb.org/siedl/). Because Web 2.0 applications are so widespread outside the library community they appear intuitively more attractive as enhancements for library web sites. This is reflected in the themes at library conferences, for example, the IFLA Conference 2008 (IFLA, 2008).

The most fertile application areas for the semantic in libraries are:

- Special collections of data with rich metadata structure which exists as, or can be automatically converted to, RDF and OWL.
- Application areas where users have commitment to annotation, tagging and sharing.
- Collaborative research groups, for example, multi-partner research teams.
- Existing physical groups, for example, local services in public libraries or readers' groups.

Research is needed on user needs and behaviour when confronted with a range of semantic web features, for example, what features do users actually need, and to what extent does the principle of least effort apply to creation of richer metadata by users. This requires installation of systems such as JeromeDL in full operational library settings with broadly based research teams consisting of librarians, computer scientists and library systems developers working together. Gradual introduction of semantic web features in ways which may be transparent to end users is more likely than a dramatic transition to new semantic web systems.

Note

1. Connotea: free online reference management for clinicians and scientists: www.connotea.org/

References

Anbarasan, E. (2000), "Tim Berners-Lee: the web's brainchild", *Unesco Courier*, September, available at: www.unesco.org/courier/2000_09/uk/dires.htm (accessed 19 August 2008).

Baker, D. (2006), "Digital library futures: a UK HE and FE perspective", *Interlending & Document Supply*, Vol. 34 No. 1, pp. 4-8.

Campbell, D.G. and Fast, K.V. (2004), "Academic libraries and the semantic web: what the future may hold for research-supporting library catalogues", *Journal of Academic Librarianship*, Vol. 30 No. 5, pp. 382-90.

Casey, M. (2005), "Working towards a definition of Library 2.0", 21 October, available at: www.librarycrunch.com/2005/10/working_towards_a_definition_o.html (accessed 10 October 2007).

Clark, K. (2005), "SPARQL: Web 2.0 meet the semantic web", 16 September, available at: www.oreillynet.com/xml/blog/2005/09/sparql_web_20_meet_the_semanti.html (accessed 10 October 2007).

Cyganiak, R. (2007), "SweolG/TaskForces/CommunityProjects/LinkingOpenData – ESW Wiki", 9 October 2007, available at: http://esw.w3.org/topic/SweolG/TaskForces/CommunityProjects/LinkingOpenData (accessed 10 October 2007).

Decker, S. (2005), "Semantic Web 2.0 … Stefan Decker on the semantic web", 11 October, available at: www.stefandecker.org/blog/archives/7-Semantic-Web-2.0 … html (accessed 15 August 2008).

Fawzi, M. (2006), "Wikipedia 3.0: the end of Google?", 3 July, available at: http://evolvingtrends.wordpress.com/2006/06/26/wikipedia-30-the-end-of-google/ (accessed 10 October 2007).

Ferran, N., Mor, E. and Minguillon, J. (2005), "Towards personalization in digital libraries through ontologies", *Library Management*, Vol. 26 No. 4, pp. 206-17.

Greenberg, J. (2007), "Advancing the semantic web via library functions", *Cataloging & Classification Quarterly*, Vol. 43 Nos 3/4, pp. 203-25.

Gruber, T. (2008), "Collective knowledge systems: where the social web meets the semantic web", *Journal of Web Semantics*, Vol. 6 No. 1, pp. 4-13.

Herman, I. (2007), "ConverterToRdf – ESW Wiki", 22 August, available at: http://esw.w3.org/toic/ConverterToRdf (accessed 10 October 2007).

IFLA (2008), "World Library and Information Congress: 74th IFLA General Conference and Council. Libraries without Borders: Navigating towards Global Understanding" 10-14 August 2008, Québec, Canada, available at: www.ifla.org/IV/ifla74/Programme2008.htm (accessed 4 March 2009).

MacManus, R. (2005), "Web 2.0 and semantic web: Mars and Venus?", 18 September 2005, available at: http://blogs.zdnet.com/web2exploer/index.php?p = 13 (accessed 18 August 2008).

Nucci, M., Barbera, M., Morbidoni, C. and Hahn, D. (2008), "A semantic web powered distributed digital library system", in Chan, L. and Mornati, S. (Eds), *ELPUB2008. Open Scholarship: Authority Community and Sustainability in the Age of Web 2.0 – Proceedings of the 12th International Conference on Electronic Publishing, Toronto, Canada, 25-27 June 2008*, pp. 130-9, available at: http://elpub.scix.net/cgi-bin/works/Show?130_elpub2008 (accessed 22 August 2008).

Zambonini, D. (2005), "Is Web 2.0 killing the semantic web", 7 October, available at: www.oreillynet.com/xml/blog/2005/10/is_web_20_killing_the_semantic.html (accessed 10 October 2007).

Corresponding author
Mary Burke can be contacted at: Mary.Burke@ucd.ie