The Semantic Revolution

Jason Bengtson

Abstract: Semantic technologies are in the process of revolutionizing the way we store, access, and communicate digital information. It is vital that information professionals be conversant in the foundational concepts upon which these technologies are based. This article uses two implementations of semantic technologies, the semantic web and the semantic medline project, to introduce the key ideas of semantic technologies to readers.

Keywords: Semantic web, Semantic technology, Summarization theory, Metadata, HTML, Graph theory, Triplestore, Semantic medline

Author Bio: Jason Bengtson, MLIS, MA is the Head of Library Computing and Information Systems at the University of Oklahoma's Robert M. Bird Health Sciences Library, 1000 Stanton L Young Blvd, Oklahoma City, OK 73117. Jason oversees the computing resources at the library and explores new technologies to determine how the library might engage with them. Jason is a member of the Association for Information Science and Technology and the South Central Chapter of the Medical Library Association. He may be reached at jason-bengtson@ouhsc.edu or through his website at www.jasonbengtson.com.

INTRODUCTION

Semantic technologies are in the process of revolutionizing the way we store, access, and communicate digital information. This article will provide readers with a basic introduction to some of the important concepts and developments in this area. Semantic technologies represent the next step in search and discovery, allowing librarians to make connections between concepts that might otherwise be missed. These technologies also represent a new level of tool for informationists in managing volumes of information which have grown too great for direct human curation, or, in some cases, direct human search. The *semantic web* offers information professionals, working in concert with their institutional Web team, tools to help users more easily find their resources, and allows software tools to collocate those resources with other, related resource portals.

THE SECOND "GREAT INVERSION"

The profession of librarianship was turned on its head with the sudden accumulation of information that occurred as a result of the widespread adoption of the printed book and, later, the digital computer. This is a transition referred to in the past by this author as *the Great Inversion*; when the role of librarian changed radically from that of caretaker to a limited amount of extant knowledge to one of a skilled organizer and navigator of the embarrassment of information riches characterized by the modern age.

Information professionals now find themselves in the midst of a second Great Inversion, as the amount of information continues to vastly outstrip the ability of those same information

professionals to manage it directly. Increasingly, a significant amount of the organization, and even discovery, of information will be carried out by computerized tools which automatically tag, index, transfer, and navigate through that information. For these tools to work properly, the underlying metadata (information describing information) upon which they rely must be correctly and judiciously applied. Tools that automate the production of metadata (as we will see with the Semantic Medline SemRep engine) must be monitored, evaluated, and optimized. Information professionals must learn to use these new search tools, leveraging them to provide extraordinary new levels of service in locating novel connections and uncovering "lost" pieces of research. The role of the information professional has never been more exciting, or more filled with the potential for re-invention.

WHAT ARE SEMANTIC TECHNOLOGIES?

In its most basic sense, ‘semantic’ refers to obtaining or possessing an understanding of something. Semantic technologies seek to provide computer software applications (“mechanistic agency”) a basic understanding of the contents of an information object, such as a Web page or an article. This allows computers to do more with this information. Through the use of semantic tools, connections between related concepts can also be more positively conveyed to computer programs.

This can be seen most clearly when discussing the ‘semantic web’. In a sense, the World Wide Web is already semantic, in that it is *human semantic*. That is to say that Web pages have been designed around the need to communicate information to human beings since the earliest days of the World Wide Web. The first Web pages were basically print documents formatted for

the computer screen and linked together through hyperlinks. As such, they were as difficult for a computer program to draw meaning from as it would be for most human beings to draw meaning from a string of ones and zeros. Just as the ones and zeros need to be translated into clear instructions for humans to begin to make sense of them, web documents need to be *marked-up* with *descriptive metadata* (information explicitly describing aspects of an information object) to help computer software make sense of their contents. However, it hasn't been until relatively recently that efforts to make Web documents more *machine semantic* have truly resonated with search providers, allowing computer programs, such as web browsers or information aggregators, to better understand the information in a web page. To a human being, there is a clear difference between:

OPEN
All weekdays from eight to five
Closed weekends

and

OPEN
Says the sign
Hanging in the bleak, dusty window
What is this cry for attention?

To a computer program, however, the second bit of bad poetry looks much the same as the first selection describing business hours. Semantic mark-up provides an invisible layer of additional description that can help computers understand that the first selection represents the business hours of the subject of a Web site. Let's look at the first set of hours again, this time with Schema.org semantic markup applied so that software programs can identify them as business hours:

OPEN

<time itemprop="openingHours" datetime="Mo-Fr 08:00-17:00">All weekdays from eight to five</time>

Closed weekends

When this is done consistently, the Web becomes a much more powerful platform. It becomes exponentially easier to, for instance, write software which finds business Web sites and aggregates their hours in one place where they may be browsed or sorted by category. This development saves human beings time, removing the need to manually search Web pages for the information they need. Web searches become far more robust when search engines are able to tell the difference between, as another example, sites that *discuss* health care, as opposed to the sites of those who *offer* health care. None of these developments are realistic, however, without formatting for Web documents that helps computers understand some of the meaning behind their contents.

WEB 3.0

An organization's Web presence is often its most visible asset. By providing and maintaining a Web site which is easy for users to connect with, as well as one which is easy for other software tools to connect to and aggregate from, libraries of all kinds can increase the visibility and value of their organizations. By understanding the ways in which the Web can potentially be leveraged for search and aggregation, librarians will be better able to search an increasingly semantic Web to separate useful information sources from those of no real value. This knowledge also increases the ability of librarians to communicate these important issues to patrons.

The semantic Web has been under development for some time, but, in many ways, it has only begun to take on a truly cohesive form in the past few years. Part of the reason for this is that the primary (and certainly most visible) way to interface with the Web has generally been through search engines. Yet, traditionally, search engines disregarded most metadata in Web pages. There were good reasons for this practice. Search engine providers were leery of Web authors packing Web pages with deceptive metadata designed to manipulate searches . . . a practice commonly carried out with keywords in the early days of the Web. In addition, the standards for the addition of more than the most basic metadata to Web pages were, and remain, fragmented.

This situation was improved upon in 2011, when the large search providers agreed to a framework for adding semantic mark-up of a limited type to Web pages[1]. This *schema.org* standard can be explored in its entirety at the schema.org Web site (<https://schema.org>). Rather than a single metadata schema (model for the structure of metadata), *schema.org* is really a collection of interoperable schemas addressing different aspects of a site.

In brief, schema.org allows descriptive metadata to be added to Web pages using a variety of metadata formats. The currently supported formats are RDFa, microdata, and JSON-LD. These formats are basically containers used to hold the metadata and position it within the Web page source code. This article will discuss the relationship between the metadata format and the schema a little later.

Schema.org itself is subdivided into separate schemas describing things such as events, organizations, people, creative works, and even actions. Individual items within a page can be marked up to indicate that they represent things like a product, a movie, an embedded video, or a recipe.

As an example, the author's Web site contains a limited amount of semantic mark-up in a couple of formats, one of which is Schema.org markup in JSON-LD. Figure 1 is an example of that mark-up. Note that in the URIs (uniform resource identifiers) below personal information has been replaced with "xxxxxxxx" for privacy considerations.

```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "Person",
  "address": {
    "@type": "PostalAddress",
    "addressLocality": "Oklahoma City",
    "addressRegion": "OK",
    "postalCode": "73117",
    "streetAddress": "1000 Stanton L Young Blvd"
  },
  "colleague": [
    "xxxxxxxxxx",
    "xxxxxxxxxx",
    "xxxxxxxxxx"
  ],
 "email": "mailto:jason-bengtson@ouhsc.edu",
  "image": "1.png",
  "jobTitle": "Assistant Professor",
  "name": "Jason Bengtson",
  "url": "http://www.jasonbengtson.com"
}
</script>
```

Legend: FIGURE 1. Schema.org metadata in JSON-LD [2]

This mark-up helps search engines properly index the pages on this site as being about "Jason Bengtson" as a person, as well as helping link the concept of "Jason Bengtson" to related topics, institutions, locations, and individuals within their indexes. It also clearly indicates to any aggregators which image on the page is an image of "Jason Bengtson".

Schema.org allows for far more extensive mark-up than what is exemplified here. There is an entire section of the Schema.org specification designed to describe medical topics and medical organizations; a fact which should be of particular interest to medical information professionals. While schema.org can be inserted in a Web page right alongside other elements, putting together the metadata and deciding where it best fits into the page takes some work.

THE METADATA STRUCTURE

Metadata Schemas

Metadata needs a few things to be truly effective. First, it needs a *schema*, which defines allowable properties and the overall structure of the data. Readers may see one such schema reflected in Schema.org's hierarchal description of "Jason Bengtson" as a *person* object, as shown in Figure 1. *@type* is particularly important in providing computer programs with information about the type of schema in use, allowing them to check the data they find in the page against the appropriate data hierarchy within *schema.org*. Without a schema to define the metadata structure, a software program could have difficulty when confronted by a Web author using a Web address or an institutional mailing code under *address*. Such a situation creates problems for software tools trying to read the metadata if they encounter mark-up that doesn't follow rules the software has been programmed to understand.

Controlled Vocabularies

Second, metadata requires a shared vocabulary (usually referred to as a *controlled vocabulary*) for many of the values in the metadata. Not all values need this (it would probably be silly to try to create a master index of human names, for instance), but for many values to be understood, or matched up to related values, they must fall within a common vocabulary. For example, the metadata above lists the author's "addressRegion" as "OK" for Oklahoma. If a software tool tries to aggregate the author into an index with other informationists from the same area, it may try to match that data up with others who share the "OK" value. However, if it does, it will probably miss those who use "Oklahoma" or "Southwest". One simple way to think about

this problem is to think of a controlled vocabulary as a schema for the values of a metadata description. This is one area where schema.org falls down a bit. While many of the values it uses don't need a controlled vocabulary, there are many others that would benefit from a controlled vocabulary's strict use. Under "MedicalCondition", for instance, while the metadata can be drilled down to a variety of properties, including "pathophysiology" and "riskFactor", there are very few of these properties in which schema.org documentation even recommends the use of controlled vocabularies to standardize the property values. One of those properties is "code", which the documentation describes as "A medical code for the entity, taken from a controlled vocabulary or ontology such as ICD-9, DiseasesDB, MeSH, SNOMED-CT, RxNorm, etc."[3]. The problems here are obvious, and, as this example should make clear, there is a significant amount of space for librarians and informationists to improve upon the state of descriptive metadata used on the Web.

Mark-up Formats

Third, while schema.org forms a logical schema for communicating descriptive metadata, the mark-up itself needs a standardized format so it can actually be inserted into a Web page. There are a number of formats in current use for this purpose. Information about these formats can be found easily by searching the site of the World Wide Web Consortium (W3C): www.w3.org. Schema.org site also contains many examples of schema.org mark-up in various formats. JSON-LD is used, which allows the description of things as objects by way of JavaScript Object Notation (JSON). RDFa, which allows relationships to be expressed in a version of the Resource Description Framework (RDF) and then added directly to the html in a Web page, is another option for supplying metadata in the web environment. Very simple

metadata can even be added to a Web page using keyword values in meta tags (Figure 2 provides an example).

```
1  <meta name="description" content="website of Jason Bengtson, MLIS, MA. This is the place to learn about my work as
   an informationist, academic, web coder, author, and researcher.">
2
3  <meta name="keywords" content="HTML,CSS,XML,JavaScript,PHP,XML,XSLT,Metadata,Information Science,Computer
   Science,Information Studies,Semantic,Complexity,Chaos,Emergent Metadata, Digital Palimpsest">
4
5  <meta name="author" content="Jason Bengtson">
```

Legend: FIGURE 2. HTML meta elements[2]

There are many standards other than schema.org that may be used for semantic Web mark-up, but given the level of schema.org adoption, as well as the simplicity of the standard, it allows for a relatively easy entry point for libraries to use semantic mark-up for themselves. As more sites begin using semantic mark-up, the World Wide Web will continue to evolve into a more useful entity, and the underlying metadata standards will continue to crystallize and become both more coherent and more flexible.

SEMANTIC MEDLINE

Unlike the World Wide Web, research databases often possess one or more consistent controlled vocabularies for their descriptive metadata. However, while this metadata provides a much richer and more accurate means of search, it often leverages only a limited amount of connectivity between related concepts. A researcher looking up information about a drug's results in treating brain tumors can build a very effective search in the Entrez tool (the National Library of Medicine Web search portal for the PubMed database collection, as well as a variety of other NCBI databases) using the Medical Subject Headings (MeSH) vocabulary. However,

this search may not reveal useful information about *other* drugs being used to treat the same kind of tumor. This is because the Entrez search tool will faithfully track down articles containing the appropriate MeSH terms, including those that fall above or below a term in the MeSH tree structure, depending on the parameters of the search. But Entrez is essentially just matching text in database fields. That is to say, the tool is not really navigating the target articles in a way designed to catch novel or lateral connections. For this, we need a different kind of technology.

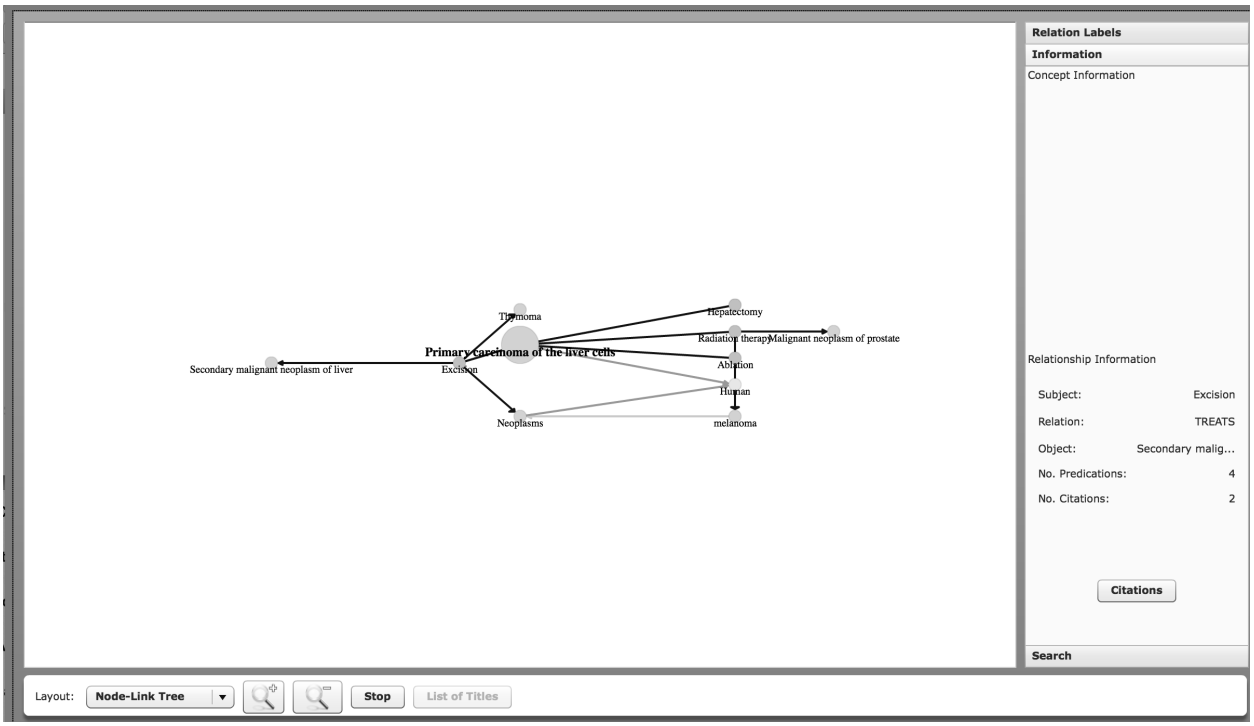
The place to look for this technology is in an alternative information structure to the tables (known technically as "relations") in a relational database. *Graph databases* organize information as *objects* and *attributes* connected by some kind of relationship (known as *predicates, arcs,* or *edges*). This basic unit of information, because it is made up of three parts, is known as a triple. A similar structure to graph databases, the *triplestore*, stores triples but does so in a relational database, or in an XML file. Table 1 is an example of a triple.

Object	Predicate/Arc	Attribute
bird_Library_Building	has_Number_Of_Floors	four

Legend: TABLE 1: A simple triple, separated into its components.

By using something like a triplestore, relationships between concepts can be traced out more completely than would otherwise be the case. This kind of concept mapping is especially useful for researchers attempting to find novel connections in the literature of their discipline. This approach also has applications for informationists navigating conceptual relationships in the hunt for search serendipity.

In a quantum leap forward for search and discovery technology, the Semantic Knowledge Representation (SKR) Project team at the Lister Hill National Center for Biomedical Communications has built, and continues to refine, a tool capable of locating and expressing these conceptual relationships between articles in Medline[4]. The *SemRep* tool uses Natural Language Processing technology (NLP) to find meaningful triples within Medline articles[5,6]. These implicit triples, once identified, are then converted by SemRep into more explicit triples containing standardized language that the *Semantic Medline* search tool can understand. This is done by mapping natural language words to terms in the National Library of Medicine's Unified Medical Language System (UMLS) Metathesaurus[6]. Relationships between terms are mapped to standardized relationships defined in a tool called the *Semantic Network*. The result is a sort of shadow database, in the form of a searchable triplestore, which may be searched and traversed using the Semantic Medline tool. All of this probably sounds a bit complicated, but the result is a visual graph that displays how concepts connect together. Figure 3 is an example of a graph created by a Semantic Medline search. In this example the arc between "Secondary malignant neoplasm of liver" and "Excision" has been clicked on to reveal information about the relationship ("TREATS") and the citations that support it. This information is visible in the panel on the right.



Legend: FIGURE 3. Screenshot of a simple Semantic Medline visual graph

In effect, we can see two types of semantic approach embodied in the Semantic Medline tool. Firstly, the tool "semantically digests" Medline articles to make them "machine semantic". By doing so, the knowledge may then be reformatted by computers in a variety of novel ways, because those computers now have an "understanding" of how the underlying concepts within the articles relate to one another. However, by creating a navigable knowledge graph for the end user, Semantic Medline also makes the knowledge within Medline more "human semantic", elucidating conceptual connections that otherwise might have completely escaped a researcher's notice.

The potential for Semantic Medline as a tool for researchers and informationists is hard to overstate. It takes search serendipity to an entirely new level of probability, and may help keep important research from languishing in obscurity . . . always a serious risk in the information age.

The Semantic Medline team has already demonstrated the visceral value of this approach. In a recent paper, Semantic Medline was used to uncover potential drug-drug interactions in existing clinical data[7]. In yet another paper, a search using Semantic Medline uncovered a possible mechanism for the so-called "obesity paradox" in critical care patients[8]. In both cases the findings were the result of new ways of navigating existing literature, rather than the completion of new research studies. The implications for information professionals are clear; increasingly, new knowledge and methodologies will be discovered by a judicious search of existing knowledge. This elevates the importance of the literature search from a contextual tool to a true research tool in and of itself, necessitating close relationships between scientific researchers and information professionals.

The Semantic Knowledge Representation project Web site contains links to the tool itself, as well as a wealth of information about the groundbreaking technologies that underlie it. This tool is currently fully functional, although it can be a little glitchy and frustrating to use at times (development on more user friendly versions continue). Information professionals may use the tool by signing up for a free account and they should expect to devote some time to reviewing the tool and its instructions. The url for the project Web site is <http://skr3.nlm.nih.gov/SemMed/>.

For those with more technical proclivities and the time to pursue them, the project has also made its triplestore downloadable for other developers to use. Full details about the structure of the triplestore are available on the project site.

The important work of the SKR project continues apace at Lister Hill. Their current research includes further refinement and development of the SemRep engine as well as further research into the application of their unique approach to literature discovery[5]. This approach may

help hospital librarians further establish value within their institutions and participate in exciting new research projects. It may also lead to a more semantic web if its underlying technologies are utilized for the automated production of metadata, as been previously theorized[9,10]. Given the improbably high resource cost of manually assigning (and continually updating) descriptive metadata to existing web sites, the automated creation and maintenance of off-site metadata indexes for search and other purposes is advisable, and entirely possible, as the technologies for automated summarization and semantic digestion continue to mature.

CONCLUSION

This article describes two significant, emerging semantic technologies as a way to introduce readers to important concepts in the area of machine semantic technology. Semantic technologies are making knowledge more discoverable, more useful, and more accessible. By leveraging semantic technologies, the Web is becoming a far more powerful platform. Researchers are beginning to see groundbreaking new technologies emerge that stand to revolutionize information seeking in general, and search and discovery in particular. These new tools are sparking a revolution in Information Science. By understanding, embracing and shaping these tools, information professionals will position themselves for the future. By helping to develop these tools, and by using these tools to leverage their Web presence, leverage Web search, and provide more powerful literature searching to users, information professionals can be a key part of that revolution.

Bibliography

1. Guha, R. “Introducing Schema.org: Search Engines Come Together for a Richer Web.” Official Google Blog, 2011. Available: <http://googleblog.blogspot.com/2011/06/introducing-schemaorg-search-engines.html>. Accessed June 22.

2. Jason Bengtson, MLIS, MA. Available:< http://www.jasonbengtson.com/>. Accessed: June 24, 2014.

3. MedicalCondition - Schema.org. Available:< http://schema.org/MedicalCondition> . Accessed: June 22, 2014.

4. Rindflesch, T.C.; Halil K.; Fiszman, M.; Rosemblat, G.; and Shin, D. “Semantic MEDLINE: An Advanced Information Management Application for Biomedicine.” *Information Services and Use* 31, no. 1 (2011): 15–21.

5. Semantic Knowledge Representation. Available: < http://skr3.nlm.nih.gov/Projects.html >. Accessed June 24, 2014.

6. Rindflesch, T.C., and Fiszman, M. “The Interaction of Domain Knowledge and Linguistic Structure in Natural Language Processing: Interpreting Hypernymic Propositions in Biomedical Text.” *Journal of Biomedical Informatics* 36, no. 6 (2003): 462–77.

7. Zhang, R.; Cairelli, M.J.; Fiszman, M.; Rosemblat, G.; Kilicoglu, H.; Rindflesch, T.C; Pakhomov, S.V.; and Melton, G.B. “Using Semantic Predications to Uncover Drug-Drug Interactions in Clinical Data.” *Journal of Biomedical Informatics* 49 (June 2014): 134–47.

8. Michael C.J.; Miller, C.M.; Fiszman, M.; T. Workman, E.T.; and Rindflesch, T.C. “Semantic MEDLINE for Discovery Browsing: Using Semantic Predications and the Literature-Based Discovery Paradigm to Elucidate a Mechanism for the Obesity Paradox.” *AMIA Annual Symposium Proceedings 2013* (November 2013): 164–73.

9. Bengtson, J. “Why I Can’t Love the Homemade Semantic Web.” 2010. *B Sides*. Available: <http://ir.uiowa.edu/bsides/20>. Accessed: June 22, 2014.

10. Bengtson, J.; Hopkins, M.; and Goodell, J. 2013. “Can Semantic MEDLINE Improve Upon the Homemade Semantic Web?” (paper presented at the annual meeting for the South Central Chapter of the Medical Library Association, Fort Worth, Texas, October 26-30, 2013).