The Semantic Web: What you Need to Know and Why it is Important

Darrell Gunter
Collexis Holdings, Inc., gunter@collexis.com

Terry Hulbert
AIP, thulbert@aip.org

Thane Kerner
Silverchair, thanek@silverchair.com

Steve Leicht
Collexis Holdings, Inc., leicht@collexis.com

Follow this and additional works at: <https://docs.lib.purdue.edu/charleston>

An indexed, print copy of the Proceedings is also available for purchase at:

<http://www.thepress.purdue.edu/series/charleston>.

You may also be interested in the new series, Charleston Insights in Library, Archival, and Information Sciences. Find out more at: <http://www.thepress.purdue.edu/series/charleston-insights-library-archival-and-information-sciences>.

Darrell Gunter, Terry Hulbert, Thane Kerner, and Steve Leicht, "The Semantic Web: What you Need to Know and Why it is Important" (2009). *Proceedings of the Charleston Library Conference*.
<http://dx.doi.org/10.5703/1288284314760>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

THE SEMANTIC WEB: WHAT YOU NEED TO KNOW AND WHY IT IS IMPORTANT FOR YOUR USER COMMUNITY.

Darrell Gunter (gunter@collexis.com) – CMO, Collexis Holdings, Inc.

Terry Hulbert (THULBERT@aip.org) - Director of Business of Development , AIP

Thane Kerner (thanek@silverchair.com) - President & Chief Executive Officer,

Steve Leicht (Leicht@collexis.com) – COO, Collexis Holdings, Inc.

Introduction

The Semantic Web: What you need to know and why it is important for your user community.

Darrell W. Gunter – EVP / CMO Collexis Holdings, Inc.

Semantic technology is now taking shape in many forms and applications in the STM industry. Understanding the difference between a Boolean search and semantic search is very important in determining what is best for your user community. Today we are going to explore the definition of semantic technology and its positive attributes and how it enhances scientific research. The focus of this session is to provide the attendees with a primer and definition of the semantic web technology and it will provide an overview of NLP and the related tools (Ontology, Taxonomy and Thesaurus) that are essential to creating a conceptual map of the data. With this foundation we will explore the semantic technology applications in the STM industry, in the commercial space and what are some of the exciting future possibilities for the semantic web. You will see applications utilizing semantic technology. Finally, we will show you how a publisher integrates semantic technology into their applications and workflow tools for the author community. After today's panel discussion you will come away with a foundation in the understanding of the semantic web technology, why it is important to your user community and what are the applications that are working today. During the session the audience will participate in a very lively Q & A with the panel.

Thane Kerner, President & Chief Executive Officer

What are Semantics and the Semantic Web?

The Semantic Web provides a common framework that allows *data* to be shared and reused across application, enterprise, and community boundaries.

--W3C Semantic Web Activity Definition

The Semantic Web requires us to go beyond documents and think of our content as data.

For example:

1 practice guideline = 1 document

OR

1 practice guideline = 312 distinct pieces of data. This comes more naturally to industries that have traditionally dealt with uniform data (finance, travel). The semantic layer is an evolution of traditional web <meta> data.

It is a consistent, rules-based information layer for computer logic parsing. It is a method for exposing the *meaning* of data so the computer can perform more sophisticated cognitive tasks.

Parallel Data

For Humans: The Narrative Layer. Take a chapter from a book; Chapter 23: Numbness, Tingling, and Sensory Loss. Normal somatic sensation reflects a continuous monitoring process, little of which reaches consciousness under ordinary conditions. By contrast, disordered sensation, particularly when experienced as painful, is alarming and...

For Computers: The Semantic Layer

```
<semantics controlvocab="UMLS">
  <tag>
    <root-term termID="28648">sensation disorders</root-term>
      <sub-term termID="180">classification</sub-term>
      <sub-term termID="6138">terminology</sub-term>
    </tag>
  <tag>
    <root-term termID="39923">sensory testing</root-term>
  </tag>
</semantics>
```

Now let's look at the infrastructure behind the order of semantic technology by complexity; Vocabularies, Taxonomies, Ontologies:

Term list - Simple set of words used in text

Controlled vocabulary - Uses only approved terms

Taxonomy - Includes structural hierarchy (parent/child)

Ontology - Limitless relationship types defined in system

The taxonomy is the framework for the semantic layer and semantic tagging—crucial for concept normalization and hierarchies. Industry standard taxonomies facilitate integration of the key concepts. Taxonomies are living creatures—they should be actively managed by an expert team (e.g. Silverchair Cortex is updated every day)

Normalization within the semantic framework is very important as authors use different terminology in different books, journal articles, and even in the same book. A semantic layer with a controlled vocabulary will normalize these differences and make user-data connections smarter. This is especially pertinent in health care.

To further refine the semantic capability there is need for further normalization for example; Synonyms (newborn = neonate), Acronyms (GHB = gamma hydroxybutyrate), Shorthand (c diff = *clostridium difficile*). *Bonus:* You can use a semantic normalization web service in your search without tagging your content.

Contextual Integration

By using a shared vocabulary or taxonomy, you can more easily integrate your varied content (journals, books, videos, images, training). Current taxonomies in health care include: MeSH, SNOMED, ICD-10, Read Codes, Silverchair Cortex, (and about 100 more). The Unified Medical Language System (UMLS) is a place to start for health care integrations. Silverchair's TOTEM Taxonomy Platform allows the user to create and maintain their very own taxonomy.

Semantic tagging is the insertion of semantic information in the XML, whose smallest unit is called a tag. Tagging can also be placed in database tables and header files if the content is inaccessible (such as images and videos). Tagging should be done at the smallest “atomic” level of data possible. Who tags and how? Human indexers are the most accurate taggers for high-value content, but computer routines can help them tag or tag extremely formulaic content. At Silverchair, we run an automated routine to place obvious tags and medical editors apply the rest. Community tagging/author tagging seems attractive, but can be risky due to inconsistency.

Immediate Benefits of Semantics. Precision in discovery! Precision in answering user queries is a key component of an application’s usability and user satisfaction rating. The semantic layer provides an application with a concise guide to the content in a language it can understand. It can now provide more accurate results. **Computable Context Links.** Create a rich matrix of contextual linking for your users using the semantic layer. These links never have to be updated by a person—semantics enable instantaneous, automated relationships whenever new content is added. **Content Intelligence!** Where are the topic gaps in your collections? Where is your content complete? Semantic reports give a unified view to integrated sites and can help guide collection development.

Trends. How are certain topics trending among your user groups? What topics are of greatest interest and value to your users?

What is the next wave of SEO? Discovery tools (intelligent agents, virtual research assistants) will give greater weight to content they can understand. Don’t let your collections be part of the “dark web”—expose your content through your semantic layer. Semantics have the potential to dramatically enhance federated search.

Ask publishers and aggregators about what semantic metadata they can provide. Many publishers are enriching content with semantic metadata now, and many more will in the future. Be sure to ask what kind of metadata is available to support your applications.

Terry Hulbert, AIP, Director of Business of Development

The significance of the semantic web and its growing importance to both publishers and users of online services in the STM industry owes much too much of the recent research related to information seeking behavior and the sheer burden of staying abreast of the research output. The number of articles read per researcher was 30% higher in 2006 than ten years earlier (1). In the same time the reading time per article has fallen, from 32 minutes in 1996, to 24 minutes in 2006. Earlier research had already shown that 98% STM researchers prefer to use online journals (2).

As a consequence of the tension between these two metrics reading patterns are changing. In a recent *Science* article it was suggested that “scientists skim journal articles to discover valuable information. They scan for terminology, segments, diagrams, and summaries of particular interest. But they don’t read individual articles left-to-right, top-to-bottom” (3). The integration and use of semantic technologies, allowing entity extraction and more, will allow this “strategic reading” and “power browsing” to be replicated in the online world as vendors move toward atomized and deconstructed articles and related content.

The implementation of these technologies allows faceting, data visualization, related linking, and more. All of these provide clues to the researcher allowing them to scan for and determine the value of an article and decide whether it merits in depth reading. As semantic technologies go

mainstream we will see the publishing community helping researchers as a consequence of improved navigation and enhanced discoverability.

- (1) C. Tenopir and D.W. King, "Electronic Journals and Changes in Scholarly Article Seeking and Reading Patterns," *D-Lib Magazine*, November/December 2008, Vol.14, No. 11/12.
- (2) B. Hemminger, D. Lu, K. Vaughan, S.J. Adams, "Information Seeking Behaviour of Academic Scientists," *J. Am. Soc. Information Sci. Technol.* 58, 2205 (2007)
- (3) Allen H. Renear and Carole L. Palmer, *Science* 325(5942), 828 (14 August 2009)

Steve Leicht - COO Collexis Holdings, Inc.

We often joke about who started the internet as Al Gore became the bunt of many jokes as he was a big promoter of the internet during his time as the Vice President in the Clinton administration. Actually Tim Berners Lee invented the [World Wide Web](#) in 1989 while working at CERN, the European Particle Physics Laboratory in Geneva, Switzerland. He wrote the first WWW client (a browser-editor running under NeXTStep) and the first WWW server along with most of the communications software, defining URLs, HTTP and HTML.

In addition to the classic "Web of documents" W3C is helping to build a technology stack to support a "Web of data," the sort of data you find in databases. The ultimate goal of the Web of data is to enable computers to do more useful work and to develop systems that can support trusted interactions over the network. The term "Semantic Web" refers to W3C's vision of the Web of linked data. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data.

The semantic web is not mature technology for global networks (yet). The current uses have more implications for Librarians than they do for web users or researchers. The power of semantic web technologies is already available today, but it is most prevalent in vertical applications with semi-structured content. Building vocabularies allows us to link data based on content is „doable‘ with many of the vertical content sources that libraries access every day or with limited integration projects. Some recent examples are; BPedia, QDOS and the University of Michigan.

Based on the current GPS position of a mobile device, DBpedia Mobile renders a map containing information about nearby locations from the DBpedia dataset. Geographic locations are currently available for 300,000 of DBpedia's 2.18 "things". It uses the Marbles Linked Data Browser to render Fresnel-based views for selected resources, as well as its SPARQL capabilities to build the map view. Starting from the map, users can explore background information about locations and can navigate into DBpedia and other interlinked datasets such as GoNames, Revyu, Eurostat and Flickr.

At the University of Michigan, they have been able to create a professional network of their entire research community utilizing the Collexis proprietary Fingerprinting technology. The Collexis Research Profile Dashboard provides the university with a view of every researcher's profile, their co-author network both within and outside of the University of Michigan. The Research Profile also shows the research of a subject, disease category, department as well as the trends in any research topic area. All of these connections are made possible by the use of semantic technology.

Semantic technology coupled with the new visualization tools provides the researcher with the exciting new knowledge discovery opportunities. Over the next 2 -3 years we will experience many new developments that will surely push the envelope of scientific research.