Key issue

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

DARRELL GUNTER
EVP/CMO
Collexis Holdings, Inc

TERRY HULBERT
Director of Business Development, American Institute of Physics

THANE KERNER
President & CEO
Collexis Holdings, Inc

STEVE LEICHT
COO
Collexis Holdings, Inc

Semantic technology is now taking shape in many forms and applications within the STM (scientific, technical and medical) industry. Understanding the difference between a Boolean search and semantic search is very important in determining what is best for your user community. We are going to explore the definition of semantic technology and its positive attributes and how it enhances scientific research. The aim is to provide the reader with a primer and definition of the semantic web technology, and it will provide an overview of Natural Language Processing (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, some of the exciting future possibilities for the Semantic Web and how a publisher integrates semantic technology into their applications and workflow tools for the author community.

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. Now let's look at the infrastructure behind the order of semantic technology by complexity; vocabularies, taxonomies and ontologies:

- term list - simple set of words used in text
- controlled vocabulary – uses only approved terms
- taxonomy (see also below) – 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 (see below) and hierarchies. Industry standard taxonomies facilitate integration

of the key concepts. Taxonomies are living creatures – they should be actively managed by an expert team.

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*). A bonus is that 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.

Tagging

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.

Further benefits of and uses for semantic technology

There are many immediate benefits of semantic technology.

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 link Publishers can create a rich matrix of contextual linking for users by means of 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 Semantic reports give a unified view to integrated sites and can help guide collection development, for example: Where are the topic gaps in your collections? Where is your content complete?

Trends How are certain topics trending among your user groups? What topics are of greatest interest and value to your users? Of course, the use of these semantic technologies is not restricted to users of content. Publishers should also be motivated to make use of these technologies in order to inform their own editorial and product development plans. Text-mining tools can identify editorial trends and patterns that can mean a publisher launching new journals in emerging areas, or even publishing more content in those existing areas where they may spot a weakness.

Discoverability; utility; strategic reading These are words and terms familiar to many STM publishers as they try to find ways to make it easier for their customers to locate precisely an article, image, graph, table or chart. No longer is the Semantic Web the domain of technical staff as they talk of Web 3.0. Business staff are now looking ever closer at the possibilities of deploying taxonomies, and how entity extraction might help them identify new areas for product development or enhance the search capabilities of existing services.

The emerging significance of semantic technologies owes much to publishers' longstanding desire to improve the granularity of their content in a way that benefits their many customers. Entity extraction, the deployment of taxonomies and more could finally allow publishers to carry out automated and deep text mining of their content and truly allow customers to precisely target only that specific content of genuine and 'real' use in their research.

Furthermore, this technology directly addresses 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 previously[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% of STM researchers prefer to use online journals[2].

As a direct consequence of the tension between these two metrics, reading patterns are changing and evolving. 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 cues 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, there is no doubt that we will see the publishing community helping researchers through the implementation of improved navigation and enhanced discoverability.

This is why all publishers should now be looking at how they might use this technology; this is precisely why 'semantic technology' is important to everyone within a publishing organization. Every publisher wants, indeed needs, their users to find their content in a timely and efficient manner. If your customers cannot easily find your content, then it may as well not exist. And if they cannot easily distil and digest the key data and information, they are very likely to move elsewhere. The sheer volume of content compels publishers to find ways that help users to search, filter, scan, annotate, and analyse fragments of content. Granted, semantic technologies are no 'silver bullet' but they can and will help to automate much of the effort required to achieve these objectives.

There are already many scientists who make extensive use of existing indexing and search and retrieval services. However, the ongoing deployment of semantic technologies and the subsequent enhancement and development of text mining, automated services, rich subject area ontologies, and so on, will help all to manage the sheer burden of keeping up to date with the seemingly endless volume of published research.

At the American Institute of Physics (AIP), semantic technologies have been used to create their recently-launched social networking service, AIP UniPHY, a collaboration with Collexis Holdings, Inc. By analysing authors and their published output, UniPHY is able to create an author profile based on the Physics & Astronomy Classification Scheme (PACS) and it is also possible to slice and view this information by affiliation and view authors within niche topic areas. However, this is only a small part of what might be possible and only demonstrates that how you might use semantic technologies is limited only by your imagination and willingness to experiment.

Although it might have taken longer than anticipated to gain traction, and many of the necessary changes will be incremental, it is becoming increasingly clear that semantic technologies are gathering momentum and we can expect to see more publishers looking at how they might use this technology in a way that helps both users and their organization.

References

1. Tenopir, C and King, D W, Electronic Journals and Changes in Scholarly Article Seeking and Reading Patterns, *D-Lib Magazine*, November/December 2008, 14 (11/12).
2. Hemminger, B, Lu, D, Vaughan,K and Adams,S J, Information Seeking Behaviour of Academic Scientists, *Journal of the American Society for Information Science and Technology, 2007*, 58, 2205.
3. Renear, A H and Palmer, C L, *Science,* 14 August 2009, 325(5942), 828.