FEATURE / MANCHETTE

Health information professionals and the Semantic Web: a symbiotic relationship?[1]

Allison McArthur

Abstract: *Objective* – To identify opportunities for health information professionals to participate in the ongoing development of the Semantic Web and to investigate the benefits promised by emerging semantic technologies. *Methods* – A search of two biomedical databases (MEDLINE, CINAHL), three information science databases (LISTA, LISA, and Library Literature), and Web-published literature retrieved articles on core Semantic Web concepts that relate to the practice of information professionals and the use of Semantic Web technologies in a health care context. Articles that focused solely on Semantic Web architecture and similar technology-focused literature were not selected for inclusion. A basic thematic analysis was performed to identify trends in the selected literature. *Results* – Five key areas of professional focus emerged as themes: ontology development, knowledge translation, information retrieval, scientific publishing, and resource classification and indexing. *Discussion* – Health information professionals have a role to play in the development of a powerful and nuanced biomedical Semantic Web. Rather than making traditional medical library positions obsolete, the Semantic Web could present new opportunities for information professionals within the five identified areas of professional focus. In turn, semantic technologies developed with the input of experienced health information professionals have the potential to transform the practice of these professionals, particularly within the five aforementioned areas.

Introduction

Although the Semantic Web is still mostly conceptual [1], it has the potential to bring about a significant paradigm shift in the field of information management, particularly with respect to medical knowledge and data. In 1998, Internet pioneer Sir Tim Berners-Lee described the Semantic Web as "a web of data, in some ways like a global database" [2]. The implications of Berners-Lee's definition are far-reaching; he effectively proposes the semantic indexing of every resource on the Web through the attachment of semantic "tags" that describe the meaning of each item in a systematic way. This set of standardized and comprehensive metadata is akin to a sort of universally applied controlled vocabulary that would make applications interoperable. Ideally, this would enable simultaneous and instantaneous querying of data from virtually any source on the Web [3]. Ultimately, the Semantic Web will make the *meaning* of Web content machine-accessible using formal rules to express the meaning of data and the relationships between concepts [4].

Despite criticisms that the notion of a Semantic Web is impractical [4], a great deal of time, funding, and attention is being devoted to the exploration of its potential. These efforts are particularly focused on developing applications in biomedicine. The Semantic Web Health Care and Life Sciences Interest Group (HCLSIG), a subcommittee of Berners-Lee's World Wide Web Consortium (W3C), has been formed to investigate the use of semantic applications for research and collaboration within the domains of health care and life sciences [5]. Leaders in the field of medical librarianship such as Dean Giustini have written at length on the subject of the Semantic Web and its implications for medical libraries [6–9].

The aim of this selected literature review is to provide an overview of promising potential uses of semantic technologies in health care and, by identifying prominent themes in published Semantic Web literature, discuss the professional areas of focus within which health information professionals can contribute to the ongoing development of the emerging Semantic Web.

Methods

Five bibliographic databases were searched to retrieve articles published in English between 1996, when the concept of a Semantic Web first began to emerge, and 2008. To locate articles that would provide an introduction to core Semantic Web concepts from the perspective of information professionals, several library and information science indices (Library, Information Science and Technology Abstracts (LISTA); Library and Information Science Abstracts (LISA); and Library Literature) were searched using a basic keyword strategy. MEDLINE and the Cumulative Index to Nursing

A. McArthur. Faculty of Information, University of Toronto, Toronto, Ontario, Canada (e-mail: allison.mcarthur@utoronto.ca).

[1]This article has been peer-reviewed.

and Allied Health Literature (CINAHL) were searched using a combination of keywords and subject headings to retrieve articles on the use of Semantic Web technologies in a health care context. Finally, a Web search was conducted to compile important grey literature pertaining to the Semantic Web, which was found on pages such as the Web site of the W3C and the blogs of prominent information specialists.

A diverse range of material was considered for inclusion in this analysis to facilitate a thorough review of the topic. Grey literature was considered in addition to peer-reviewed journal articles to allow for inclusion of seminal Semantic Web materials that were not published via formal channels. Literature that focused solely on Semantic Web architecture was not included.

Included studies were analyzed by themes derived from the literature. Each article was coded according to the professional areas of focus (related to the practice of health information management) that it touched upon. Relevant sections of articles were extracted and grouped according to emerging themes.

Results

This review revealed five key areas of professional focus within which health information professionals may shape the evolution of the Semantic Web: ontology development, knowledge translation, information retrieval, scientific publishing, and resource classification and indexing. In turn, semantic technologies also have the potential to streamline and enhance professional practice in these areas. Should widespread development of semantic applications occur in these areas, health information professionals will be in a favourable position to utilize and market their unique skill set.

Discussion

Ontology development

Though the Semantic Web will be made possible by new and complex innovations in information technology, the massive universal ontologies enabled by these innovations will be developed using established classification techniques. Ontologies are critical to a successful Semantic Web [10]. Cho and Giustini [8] observe that the Semantic Web will require the classification of billions of resources, not unlike what Melville Dewey did for print materials. The ongoing process of ontology development, integration, and maintenance is fundamental to the success of the Semantic Web and will provide opportunities for librarians to make meaningful contributions to its development. Health information professionals have valuable experience dealing with the proliferation of polysemy (the ambiguity of a term that has two or more meanings depending on context) in medical terminology, which is sure to complicate the development of ontologies for medical information. Medical polysemy has the potential to impede natural language processing, interfere with the definition of terminological standards, and generally hinder intelligent information access, all of which are crucial to the diffusion of the Semantic Web [11].

Robu et al. [4] assert that "the importance of medical librarians in this process is not only unlikely to decrease, but their role will become even more crucial. Developing the various ontologies and medical taxonomies cannot lead to any useful real-life applications without major input from librarians." The expertise of health information professionals has great practical value; it could help to transform the Semantic Web from a lofty vision into an everyday reality.

Resource classification and indexing

Indexing is a highly specialized skill possessed by information professionals, most of whom must apply it in some form in their day-to-day work. Given the exponential growth of the body of published biomedical literature [12], health information indexers are in high demand. It seems inevitable that this demand will be multiplied yet again should the vision of a biomedical Semantic Web be realized. Once an ontology has been developed, resources must be indexed according to the concepts and relationships it defines. The specialized skill set required for quality detailed indexing and the time-consuming nature of the process will increase the need for experienced indexers.

Semantic Web indexing will also be complicated by the fact that ontologies will necessarily change as language evolves, which will necessitate periodic re-indexing [13]. In anticipation of this inflating demand, some researchers are endeavoring to use natural language processing to perform automated indexing. One such automated indexing project is the US National Library of Medicine's Indexing Initiative (IND), which explores ways to partially or completely substitute current indexing practices with automated indexing [14].

This and other automated systems are still in their infancy. Ferguson [13] notes that this type of processing will not be useful for tagging deep etymological concepts, and to train algorithms for automated indexing software, immense control sets of consistently indexed documents will be needed. Extraction of concepts from unstructured publications such as biomedical research articles is notoriously difficult to accomplish using automatic indexing tools [4]. It is also important to note that these tools are not useful for indexing materials that are not text-based, such as medical images, diagrams, videos of operative procedures, and other specialized medical resources [15].

If automated indexing tools continue to be tested and refined, they may eventually make the task of indexing easier and more efficient for health information professionals, much like the benchmarking tools and quality checklists currently in use. However, they will never match the sensitivity and precision of manual indexing, a fact that is stressed even by the manufacturers of indexing tools [15].

Information retrieval

Advanced information retrieval, a core skill of health information professionals, will be simplified by a biomedical Semantic Web. The most important benefit of Semantic Web technologies for medical librarians is that they will enable the so-called semantic search. These technologies will allow users to specify the precise meaning of ambiguous terms in a query, such as "cold" (temperature or disease) [1]. The Semantic Web may ultimately eliminate the need to search multiple databases with irreconcilable subject headings by linking and augmenting (but not replacing) biomedical databases and making them more accessible, thereby facilitating

collaboration across knowledge domains and streamlining the process of information retrieval [13].

The Semantic Web may reduce the need to purchase fee-based databases, as it will be equipped with an equivalent or superior level of behind-the-scenes concept mapping and relationship definition, although it remains to be seen who will fund and carry out this mapping. The volume of information that is available will remain the same in a Semantic Web environment, but it will be more effectively classified. Health information specialists will also have more specialized Semantic Web searching tools available to them that, in theory, will make information retrieval more efficient. HealthCyberMap is a beta version of one such tool, which is designed to map online health information resources in new semantic ways to simplify end user navigation and retrieval "through intelligent categorization and interactive hypermedia visualization of the health information cyberspace" [15]. Though these tools promise increased usability, health information professionals will continue to be experts in using and troubleshooting semantic search engines and databases. Ideally, health information specialists will take the opportunity to pilot test these types of search tools in the design phase and provide feedback to developers.

Knowledge translation

Effective knowledge translation (KT) within and among health-related organizations is a challenging task. The many subspecialties of medicine, each with its own lexicon, pose challenges to effective KT as a result of rapid and efficient innovation that produces subcultures that do not speak the same language. The concept of a unique resource identifier (URI), proposed as part of the Semantic Web, would address this issue by allowing users to effortlessly communicate newly invented concepts. The use of URIs would enable software agents to analyze, interpret, and translate human expression in such a way that the knowledge created within a particular domain could be understood by anyone [16]. A biomedical Semantic Web holds the potential to facilitate the KT process by translating highly specialized medical knowledge into universally understandable concepts via semantic metadata.

Semantic Web technologies promote both the sharing and understanding of information across diverse domains of expertise. They simplify and standardize the definition of concepts, allowing widespread access not only to the physical form of information (journal articles, datasets) but also to something less tangible: an informed understanding of the concepts contained in those physical representations. The capability of the Semantic Web to compile information from diverse sources also allows researchers to immediately and collaboratively annotate scientific concepts based on their insights, newly proposed hypotheses, or the disproof of a theory [3]. Successful implementation of a biomedical Semantic Web would encourage the active participation of researchers in cross-disciplinary KT.

The role of health information professionals in supporting KT will have to adapt to support these emerging developments. Among their key objectives will be educating their clientele about these new KT processes and mechanisms and developing systems and applications that maximize the benefits of these new Semantic Web technologies, as information professionals once did for now-conventional Web capabilities.

Scientific publishing

The scholarly publishing cycle may soon be expedited and reconfigured by the capabilities of Semantic Web technologies. As open access publishing gains momentum, authors will be able to add machine-readable semantic metadata to their own papers using specialized tools for Web publishing, and this metadata will allow their papers to be retrieved by powerful semantic search engines [17]. Biomedical ontologies will allow researchers to increase the accessibility (and the potential impact) of their work by defining the sophisticated and systematic metadata that they will apply to their articles.

In a Semantic Web environment, researchers may be further encouraged to share their results with peers before they are formally published. Berners-Lee and Hendler [17] predict that it will be easy to find out about studies that are in progress and to view the results of these studies well before they are published in a research paper, which will in turn support less formal but more frequent discourse to inform the research process. Though this shift will create challenges for publishers, it will provide significant advantages to individual researchers and the progress of scientific innovation as a whole. Enhanced information tools will be able to capitalize on this pre-publication data by collecting and combining it with related data in ways that provide added value and insight but also preserve its original meaning. The ability to create and manipulate this "recombinant data" empowers researchers to maximize the utilization of their data [3]. The age of the Semantic Web could bring about unprecedented levels of uptake and application of research findings. Health information professionals must be able to assist their clients in annotating, disseminating, recombining, and utilizing data in new ways.

Conclusions

Robu et al. suggest that "the underlying problems currently faced in Semantic Web research have been studied for years by librarians, long before the emergence of the Web itself" [4]. Within the areas of professional focus that I have identified, health information professionals may find opportunities to apply their knowledge of classification systems, controlled vocabularies, metadata, knowledge management, and the cycle of biomedical scientific publishing to the development and use of semantic applications. It is within these areas that Semantic Web technologies developed with the input of experienced health information professionals have the potential to transform information management practices. The roles of information professionals will undoubtedly change in a Semantic Web information environment, but proactive and adaptive health information professionals are in a position to make valuable contributions to the development of semantic applications and influence the future of health information.

References

1. Mukherjea S. Information retrieval and knowledge discovery utilising a biomedical Semantic Web. *Brief Bioinform*. 2005 Sep;6(3):252–62.
2. Berners-Lee T. *Semantic Web road map* [monograph on the Internet]. Cambridge (MA): World Wide Web Consortium; 1998 [cited 2009 Apr 2]. Available from: http://www.w3.org/DesignIssues/Semantic.html.
3. Neumann E, Prusak L. Knowledge networks in the age of the Semantic Web. *Brief Bioinform*. 2007 May;8(3):141–9.
4. Robu I, Robu V, Thirion B. An introduction to the Semantic Web for health sciences librarians. *J Med Libr Assoc*. 2006 Apr;94(2):198–205.
5. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, et al. Advancing translational research with the Semantic Web. *BMC Bioinformatics*. 2007 May 9;8 Suppl 3:S2.
6. Giustini D. Web 3.0 and medicine: make way for the semantic web. *BMJ*. 2007 Dec 22;335(7633):1273–4. Available from: http://www.bmj.com/cgi/reprint/335/7633/1273.pdf.
7. Cho A, Giustini D. Web 3.0 and librarians. *J Can Health Libr Assoc*. 2008 Mar;29:13–18. Available from: http://article.pubs.nrc-cnrc.gc.ca/RPAS/rpv?hm=HInit&calyLang=eng&journal=jchla&volume=29&afpf=c07-035.pdf.
8. Cho A, Giustini D. The Semantic Web as a large, searchable catalogue: a librarian's perspective. *Semantic Universe*. 2007. Available at: http://www.semanticuniverse.com/articles-semantic-web-large-searchable-catalogue-librarian%E2%80%99s-perspective.html.
9. Giustini D. The Semantic Web is about making connections. In: The Search Principle Blog. Vancouver (BC): University of British Columbia; 2008 Mar 12 [cited 2008 Apr 6]. Available from: http://blogs.ubc.ca/googlescholar/2008/03/the-semantic-web-is-about-making-connections/.
10. Nardon FB, Moura LA. Knowledge sharing and information integration in healthcare using ontologies and deductive databases. *Medinfo*. 2004;11(Pt 1):62–6.
11. Pisanelli DM, Gangemi A, Battaglia M, Catenacci C. Coping with medical polysemy in the Semantic Web: the role of ontologies. *Medinfo*. 2004;11(Pt 1):416–9.
12. Straus SE, Sackett DL. Using research findings in clinical practice. *BMJ*. 1998 Aug 1;317(7154):339–42.
13. Ferguson JC. Semantic Web technologies: opportunity for domain targeted libraries? *J Electron Resour Med Libr.* 2007;4(1–2):113–25.
14. Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, Rindflesch TC, Wilbur WJ. The NLM Indexing Initiative. *Proc AMIA Symp*. 2000;17–21.
15. Boulos MN. A first look at HealthCyberMap medical semantic subject search engine. *Technol Health Care*. 2004;12(1):33–41.
16. Berners-Lee T, Hendler J, Lassila O. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Sci Am*. 2001. Available at: http://www.sciam.com/article.cfm?id=the-semantic-web.
17. Berners-Lee T, Hendler J. Publishing on the Semantic Web. *Nature*. 2001 Apr 26;410(6832):1023–4.