The Semantic Web, Universalist Ambition and Some Lessons from Librarianship

Terrence A. Brooks
The Information School, University of Washington
Box 352840, Seattle, WA 98195-2840
(voice) 206 543-2646
tabrooks@u.washington.edu

Abstract

Building the semantic web encounters problems similar to building large bibliographic systems. The experience of librarianship in controlling large, heterogeneous collections of bibliographic data suggests that the real obstacles facing a semantic web will be logical and textual, not mechanical. Three issues are explored in this essay: development of a standard container of information, desirability of standardizing the information hosted by this standardized container, and auxiliary tools to aid users find information. Value spaces are suggested as a solution.

A Vision of Shared Meaning

Increasing the intelligibility of the Web is a compelling vision. Imagine how the utility of local data could be enhanced if they were meaningfully linked to data posted by strangers far away. The Web could evolve into a comprehensive meaning system, a universal encyclopedia or "world brain," as prophesized by H.G. Wells (1938). Clever programs could roam this meaning space discovering useful, unanticipated information, emulating Bachman's (1973) vision of database programmers navigating an n-dimensional database space.

The extensible markup language (XML) and its attendant technologies is the fundamental facilitator of the semantic web (Berners-Lee, 2001). XML replaces presentation markup, e.g.: *<H1> My name is Terry </H1>* with markup that provides a context for understanding the meaning of the data, e.g.: *<name> Terry </name>*. Cagle (October 26, 1999) describes the era of the "distributed object" where XML elements would "roam the Internet as autonomous units in a sea of contextual relationships." Semantic markup could be potentially exploited in many ways; for example, disambiguating information resources and aiding information discovery in a rapidly expanding and heterogeneous Web. Problems like the following could be solved:

> In addition, this markup makes it much easier to develop programs that can tackle complicated questions whose answers do not reside on a single Web page. Suppose you wish to find the Ms. Cook you met at a trade conference last year. You don't remember her first name, but you remember that she worked for one of your clients and that her son was a student at your alma mater. An intelligent search program can sift through all the pages of people whose name is "Cook" (sidestepping all the pages relating to cooks, cooking, the Cook Islands and so forth), find the ones that mention working for a company that's on your list of clients and follow links to Web pages of their children to track down if any are in school at the right place. (Berners-Lee 2001)

The "Ms. Cook" Retrieval Problem

Finding a particular "Ms. Cook" in a semantic web is essentially an information retrieval problem, similar to the bibliographic problem of finding an author named "Ms. Cook." Librarians possess considerable experience dealing with this sort of problem. Their strategy for controlling bibliographic data can be summed up in a few words: Make the structural form of the data predictable, make the information contents hosted by this form predictable, and where information is structured arbitrarily, provide access tools to help the searcher find the difficult-to-anticipate information.

In some ways a semantic web and large bibliographic databases are similar. A semantic web is a single meaning system organizing a large collection of widely disparate information. So are large bibliographic databases. For example, the WorldCat database (sponsored by OCLC, Online Computer Library Center at http://oclc.org/home/) is a union catalog that hosts about 41 million records (as of Spring 2001) in 400 languages and indexes a heterogeneous collection of material including books, maps, films and slides, sound recordings, and so on. The WorldCat database has been called the most important database in academe (Smith, 1996).

A semantic web and large bibliographic databases also both employ expressive data structures. The Machine Readable Cataloging (MARC 21) record provides each field and subfield with a semantically significant field number or code. Usage conventions define exactly what sort of data can be placed in each field and subfield. One can distinguish, consequently, "John F. Kennedy" as the author of a work, the subject of a work, a person named in the work, and so on. XML also permits the definition of element names that express usage aspects of a

personal name; for example, one can create tags such as <*author*>, <*subject*> or <*named person*>.

There are, of course, great systematic differences between the Web and large bibliographic databases. The Web is magnitudes larger. It is growing faster. The origins of Web pages are not a few cooperating agencies. Web pages do not reflect a single, well-groomed record structure. Web pages do not benefit from coordinated activity distinguishing the author "Ms. Cook" from "Ms. Cook" as the subject of XML data. Furthermore, Web pages have no coordinated activity distinguishing "Mary Cook" from "Sally Cook," or even this "Mary Cook" from that "Mary Cook." This problem is commonly encountered when one uses a Web tool to search for "Mary Cook," receives hundreds of thousands of Web pages in return, and discovers that the vast majority are irrelevant.

Librarians have been struggling with these problems for decades. It is possible that their practical experience dealing with bibliographic data could be profitably applied to the semantic web proposal, especially if an exemplary semantic web activity were searching for a certain "Ms. Cook" in a heterogeneous, rapidly growing and decentralized Web.

Principal Elements of A Bibliographic System

The basic strategy for constructing a bibliographic database system is standardizing the container of the information, structuring the information contents within this container, and then building ancillary tools that aid the anticipation of the user.

The following are example methodologies and technologies:

- *Standardize the container of information*. The library community has cooperated in developing MARC records and agreeing on the usage of its fields, subfields and indicators. An example is the 700 Added Entry—Personal Name field (a description is available at http://www.oclc.org/oclc/bib/700.htm). Subfields give elaborating information such as titles and dates associated with the name, even a fuller form of the name. In Spring 2001, two "Mrs. Cook" were listed in the WorldCat database (there is no Ms. Cook listed):

  #aCook, #c Mrs., #d fl.1735-1740, #e bookseller

  #aCook, #c Mrs., #d d. 1826

We can infer that these are two different Mrs. Cook based on the *d* subfield ("Dates associated with a name") and the *e* subfield ("Relator term"). Without these auxiliary, contextual subfields, these two different Mrs. Cook could easily be misconstrued as the same person.

It is likely that in a semantic web, which lacked an agreement about supplying qualifying information, these two Mrs. Cook would have been mistakenly conflated to one person.

- *Formalize the construction of information.* The construction of the majority of the fields of the MARC record is controlled by tools such as the Anglo-American Cataloguing Rules (Gorman, 1998) and Library of Congress Subject Headings (Cataloging, 2001). Special rules exist for the construction of a surname, for example:

  > "22.15A If the name by which a person is known consists only of a surname, add the word or phrase associated with the name in works by the person or in reference sources." Example: Read, Miss (Gorman, 1998, p.410)
  >
  > "22.15B1 Add the term of address of a married woman if she is identified only by her husband's name." Example: Ward, Mrs. Humphry (Gorman, 1998, p. 410)

  Examples of the application of these rules appear above where both records construct Mrs. Cook's name as *#aCook, #c Mrs.*

  Semantic web discussion has yet to broach this deeper level of standardization. It highly probable that several XML sources may have similar <*name*> elements, and may be referring to the same person, but a robot spider would be stymied recognizing the equivalence of "Sally Cook," "Cook, Sally," "S. Cook," and "Cook, S" or any other of the innumerable variations possible in the construction of a person's name.

- *Aid Users' Anticipation.* It has been widely recognized that there are many names for the same thing (Furnas, 1983). Semantic dispersion (one person having multiple names) and semantic conflation (many people sharing the same name) are typical problems of bibliographic systems. Librarians have developed name

authority files to ensure that a single bibliographic reference points to the same person. For example, Captain James Cook (1728-1779) has references from alternate spellings and renderings including: "James Cooke," "Dzhames Kuk," "Hakobos Gowg," and "Jacques Cook" in the WorldDat database. Name authority files aid users by leading them from "Dzhames Kuk" to "James Cook." This is so powerful and efficient method of finding information that some bibliographic systems are designed for searchers to navigate authority files before jumping into the bibliographic database.

The semantic web proposal suggests the evolution of the World Wide Web into a single meaning system. Therefore, it is possible that the crucial information element about the targeted Ms. Cook is associated with a <*name*> element containing "Ms. Kuk," or perhaps "Ms. Gowg," or something else I can't anticipate. As a general rule in bibliographic systems, if you can't anticipate the representation of the information your seeking, you're going to have a hard time finding it, and so will your robot spider.

Decentralization and Its Effect on Meaning

Comparing the semantic web proposal and bibliographic databases illustrates the difference between open and closed information systems. Closed systems can impose standards on information structure and content not possible in open systems. A semantic web would be an open system, its *raison d'etre* is to find meaningful data posted by strangers far away; in short, a semantic web has universalist ambitions, yet will operate in an open environment.

XML namespaces (http://www.w3.org/TR/REC-xml-names/) addresses some of the semantic conflation problem that would exist in a semantic web. Namespaces are useful when an XML document pulls data from several XML sources and finds element name collisions. For example, a relatively common XML element like <*dollar*> could be disambiguated by reference to one namespace that contexts it as a U.S. dollar amount and another namespace that contexts it as a Canadian dollar amount.

XML namespaces do not solve the deeper semantic problem, however, that precise agreement about the meaning of any common word is rare. "Price," "revenue," "assets" and so on, can have multiple connotations depending on context. A spider robot could find many XML

sources with <*heaven*> as an element name, but do the authors mean the same thing by this term?

> Consider this illustration. At a summit of religious leaders, aimed at increasing common understanding among the world's religions, it is decided that everyone will speak English and use the vocabulary of Protestant Christianity. But as soon as the discussions start, there are problems. Someone uses the word *heaven* and many people nod in recognition. But as the discussion progresses, it is clear that even the different Christian delegates have understood different nuances of the word, let alone the Hindu and Buddhist representatives. As time goes by, they realize that perhaps they should have agreed at the start not to use a single vocabulary but rather to describe what the relationships were between the apparently similar words in the vocabularies with which they were already familiar. (Phipps 1999)

XML schemas (http://www.w3.org/TR/xmlschema-1/) formalize the syntax and value constraints of XML instances, and facilitate the sharing of information among communities of users. Biztalk.org (http://www.biztalk.org/home/default.asp) and XML.org (http://www.xml.org/) are examples of registries for schemas. The rapid development of schemas can be viewed as a positive trend for the penetration of XML, but an unintended consequence is a growing lack of transference among schemas:

> As the pace of activity around the Web-based XML schema repositories accelerates, the number of registered schemas increases dramatically. As an example, a quick count revealed that among the more than 300 schemas at one of the major XML repositories, at least eight of these describe purchase order documents. As a developer, which should I choose? Regardless of my choice, if my company deals with multiple trading partners—and what company doesn't—which schemas will my partners choose? And across these multiple schemas, how many different tag names for "customer number," "ship-to address," or "purchase order number" are there like to be? (Lewis 2001)

The only solution to schema proliferation is the convergence on a few schemas that will act as touchstones or translation devices for a community of users. This emulates the development of the MARC record structure as a common structuring device for bibliographic data. Even though parochial MARC formats exist; for example, Canadians have CANMARC,

Finns have FINNMARC, Hungarians has HUNMARC, the library community shares data uniformly structured as MARC records.

The Universal Data Element Framework (UDEF) at http://www.udef.com/, which describes itself as a "Dewey Decimal-Like Indexing System" for the Web, is a possible device for rationalizing the tags of XML schemas. UDEF would index schema tags as they are submitted to registries and, if the system were ever to be widely implemented, would supply semantically equal tags across multiple interest domains.

Proposals such as UDEF represent centrist impulses that are at odds with the open, unregulated nature of the Web. Success of the UDEF would depend on the cooperation of the Web community, a doubtful prospect at best. Partial deployment of the UDEF solution suggests a partitioned semantic web where clusters of Web pages would become intelligible through the translation device of one or more tag indexes, while other Web pages posted by individuals, or organizations that refuse to participate would be missing. A partial semantic web that systematically missed whole classes of Web pages doesn't seem to manifest the spirit of a decentralized, yet single-meaning worldwide system.

The Impediment of Orthography

The experience of librarianship in organizing large depositories of bibliographic information suggests that the success of a semantic web hinges on the deep standardization of the information content of XML elements. For example, a spider robot may successfully locate a <*name*> element, but yet be stymied matching "Hakobos Gowg" to "James Cook."

Orthography is uncontrolled on the Web: There is no worldwide law enforcing the use of language. Variant spellings, contractions, neologisms have proven a fundamental impediment to online information retrieval (Brooks 1998). The problem is so well recognized that commercial database vendors construct databases specifically to help users with the scatter of company names (i.e., DIALOG Company Name Finder database), products (i.e., DIALOG Product Code Finder) and journal names (i.e., DIALOG Journal Name Finder). Consider the problem of finding XML information associated with the journal "Scientific American." The DIALOG Journal Name Finder database reveals that this journal has been represented in many different ways. Here is a sample:

SCI AM
SCI AM (NEW YORK)
SCI AM NEW YORK
SCI AM.
SCI AMER
SCI. AM. (INT. ED.) (USA)
SCI. AM. (USA)
SCI. AMERICAN
SCI. AMIC.
SCI., AM.
SCI.AM
SCIENTIFIC AM
SCIENTIFIC AMERCIAN
SCIENTIFIC AMERICA

DIALOG File 414, May 8, 2001

Clever programs could be constructed to deal with these few English-language variants for one journal name, but this solution doesn't scale up if one has to anticipate the spelling variants, abbreviations and punctuation irregularities in, say, the 400 languages represented in the WorldCat database.

Uncontrolled orthography has already played an important part in the Napster controversy, and illustrates that successful sharing or retrieval of information hinges on being able to anticipate its construction:

> Which of the following is the correct spelling of the 1962 nonsense surf classic by the Rivingtons? Is it a) "Pa Pa Ooh Mow Mow," as one authoritative record collectors' Web site has it; b) "Papa Oom Mow Mow," as claimed by another; or c) the socialistically correct "Poppa Ooh Mao Mao," as avowed by a third?… But, given the court's ruling that Napster must now find a way to identify and block copyright-infringing song files, the proper spelling of song titles will likely be more than a fanboy's parlor trick: it could be a key in determining the future of Napster. (Mann 2001)

A Suggestion: Valuing Authority First

The Western Library Network architecture (WLN 1993) encouraged users to begin searching authority files before the bibliographic file. Using the "Ms. Cook" example, one could sort through the small number of authority records that established the various "Ms. Cook" and find the one you're seeking. The textual problem of your information need was solved: You knew what your target information looked like. Armed with this information, one could apply it to the bibliographic file and specify that the information resided in a certain field such as author, subject or named person. This solved the logical problem of specifying the role that your target name held in relation to the information you were seeking. Essentially, you knew what the information looked like and where it should reside.

Extension of this idea to the semantic web suggests the development of a "value space," akin to namespaces. Perhaps XML sources could link to a value space where a spider robot could find information collating "Ms. Cook" to "Mary Cook," "M. Cook," "Cook, Mary" and so on. This would, of course, add overhead to XML sources and might be more conformity possible in an unruly World Wide Web. On the other hand, there may be a central value depository, a sort of worldwide name authority file. Visiting this central value depository, I could sort through the relatively few Ms. Cook entries and select the correct individual. I could then program my spider robot to look for this formulation of the name in specific XML elements. Armed with the target formulation of the data, and knowing which XML elements to examine, my spider robot would meet with much more success because its task would have been reduced to the mechanical one of simply looking for matches.

Conclusion

> The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation. (Berners-Lee 2001)

The semantic web proposal has an unsolved tension between a universalist ambition and the need to centralize to support intelligibility. The experience of librarianship in building large bibliographic systems suggests that standardization is the key to success.

References:

[Bachman 1973] Charles W. Bachman. "The Programmer as Navigator" Communications of the ACM, v. 16 (11), 653-658.

[Berners-Lee 2001] Tim Berners-Lee, James Hendler & Ora Lassila. The Semantic Web. Scientific American, April 11, 2001[Online: http://www.sciam.com/2001/0501issue/0501berners-lee.html

[Brooks 1998] T. A. Brooks. "Orthography as a Fundamental Impediment to Online Information Retrieval." Journal of the American Society for Information Science, v. 49(8), 731-741. [Online: http://faculty.washington.edu/tabrooks/Documents/Ortho3.html]

[Cagle October 26, 1999]. Kurt Cagle. "Why XML? A look at XML and how it will change the world." [Online: http://209.20.234.180/sig/OctMtg/Presentation.xml]

[Cataloging 2001] Cataloging distribution service. "Tools for authority control-subject headings" [Online: http://www.loc.gov/cds/lcsh.html]

[Furnas 1983] G. W. Furnas, T. K. Landauer, L. M. Gomez, & S. T. Dumais. "Statistical semantics: Analysis of the potential performance of key-word information systems." The Bell System Technical Journal, 62,1753-1806.

[Gorman 1998] Anglo-American Cataloguing Rules, 2d ed. Edited by Michael Gorman and Paul W. Winkler. Chicago, IL: American Library Association.

[Lewis 2001] William J. Lewis. "XML microstandards" [Online: http://www.intelligententerprise.com/000428/supplychain.shtml]

[Mann 2001] Charles C. Mann. "Napster Will Remove Copyright-Protected Songs" [Online: http://www.thestandard.com/article/0,1902,22189,00.html]

[MARC 21] The MARC 21 Formats: Background and principles. [Online: http://lcweb.loc.gov/marc/96principl.html]

[Phipps 1999] Simon Phipps. "Meaning, not markup" [Online: http://www-106.ibm.com/developerworks/library/meaning.html]

[Smith 1996] K. Wayne Smith "OCLC - Moving toward the next stage of the electronic library." In the Proceedings of the Fourteenth Annual Conference of Research Library Directors. Tomorrow's Access-Today's Decisions: Ensuring Access to Today's Electronic Resources (pp. 1-5). Dublin, OH: OCLC Online Computer Library Center.

[Wells 1938] H. G. Wells. World Brain. Garden City, NY: Doubleday, Doran, 1938.

[WLN 1993] WLN online searching manual. Lacey, WA: WLN, 1993.