[Information Research, Volume 6 No. 2 January 2001]
Where is meaning when form is gone? Knowledge representation on the Web, by Terrence A. Brooks
Location: http://InformationR.net/ir/6-2/paper93.html

Last updated: 21st December 2000

Where is meaning when form is gone? Knowledge representation on the Web

Terrence A. Brooks
The Information School
The University of Washington
Seattle, WA 98195

Abstract

This essay argues that legacy methods of knowledge represenation do not transfer well to a Web environment. Legacy methods assume discrete documents that persist through time. Web documents are often products of dynamic scripts, database manipulations and caching or distributed processing. The size and rate of growth of the Web prohibits labor-intensive methods such as manual cataloguing. This essay suggests that an appropriate future home of content-bearing metadata is extensible markup technologies. Meaning can be incorporated in Extensible Markup Language (XML) various ways such as semanticaly rich markup tags, attributes and links among XML sources.

Introduction

How shall knowledge be represented on the Web?

Legacy knowledge representation methods assume that information takes document-like embodiments (Svenonius, 2000, p.8). The classic example occurs when a librarian tags a book with a subject heading. How well does this document paradigm transfer to the Web? Are some of its assumptions at risk in the Web environment?, i.e., *Can we treat Web pages like books? Is it economically possible to catalogue the Web?* Have extensible Web technologies for marking up resources antiquated the document as a vessel of information? If we abandon the document paradigm, where shall we locate our signifiers of meaning such as subject headings, or the Web equivalent, meaning-bearing metadata?

This essay suggests that new locales of meaning in extensible technologies may be:

- The structure of the information resource itself, implicitly in the markup tags, and explicitly as meaning-bearing attribute qualifiers
- The relationships among information resources: implicitly in the links, and explicitly as meaning-bearing attribute qualifiers
- Situational expertise that orients information seekers to the semantic norms of a specific community of information users

Legacy Methods of Knowledge Representation

Legacy methods of knowledge representation represent our starting point for handling the conceptual and technical challenges of the Web. Like any methodological practice, they reflect their technological origins and environment. In the 20th Century, typical knowledge representation technologies included the Machine Readable cataloguing (MARC) record structure, the Library of Congress Subject Headings (LCSH) and the WorldCat database. These

technologies reflect a historic transition from paper-based to digital systems, and express assumptions such as:

- **A uniform catalogue entry can represent holdings** The MARC record acts as a uniform record structure for the representation of disparate information types, including manuscripts, archives, cartographic material, musical scores, serials, sound recordings and so on. MARC is an "integrated format defined for the identification and description of different forms of bibliographic material." (The MARC 21 formats)
- **A single list of subject terms can provide subject access.** The LCSH acts as a general list of subject terms and phrases. "As an increasing number of other libraries have adopted the Library of Congress subject headings system, it has become a tool for subject indexing of library catalogs in general." (Library of Congress Subject Headings)
- **A single database can store information.** The WorldCat database acts as a single repository for information. As of June 2000, WorldCat hosted 40 million MARC records. It "is the most consulted database in academe." (Smith, 1996: 1)

These three technologies are merely representative. Other dominant 20th Century information providers such the Dialog Corporation reflect similar assumptions. The Dialog Corporation vends access to approximately 500 databases, each of which may have a unique record structure, and a list of subject terms or descriptors. In the early days of online information retrieval, it was common to refer to specific databases as "the medical database" (i.e., MEDLINE, Dialog database 154) or "the education database" (i.e., ERIC, Dialog database 1).

The conceptual and technological legacy methods of knowledge representation reflect an era when information oligarchs amassed large, unique databases to which they vended access. Consequently, some of the assumptions of knowledge representation that we carry forward to the Web are:

- A single, multi-purpose record structure may be sufficient
- Database records persist through time and will not disappear or transform into something else
- There are information professionals who develop and employ subject terms and phrases
- Aggregating information into a few large databases is useful and efficient

The Knowledge Representation Environment of the World Wide Web

Since the introduction of the Hypertext Markup Language (HTML) in 1990, the World Wide Web has become a major information utility, and will probably be the dominant paradigm for knowledge representation methodologies in the future. Can legacy knowledge representation methods be smoothly shifted to the Web?

Theoretically at least, the Web permits anyone anywhere to post pages on any topic and in any language. Such extreme decentralization makes estimating the Web's size and rate of growth difficult. A survey by Lawrence and Giles (1999) estimated 800 million public Web pages available in February 1999. In June 2000 the search engine Google claimed an index of 1 billion URLs (Google, 2000). Jacob Nielsen (1995) suggests that the growth rate of the Internet is 100 percent per year. Whatever the exact figures may be, the Web is a large, heterogeneous, decentralized phenomenon with a high rate of growth.

> So does this mean Google is the first search engine to give 100 percent coverage of the web? No. For one thing, that 1 billion page estimate is several months old, and the web has almost certainly increased in size since then. Nor does that estimate include the millions of pages that search engines typically don't crawl, such as those behind password protected areas or served up by identifiable dynamic delivery systems. How big the web is now is anyone's guess. (Sullivan, July 5, 2000)

Some parts of the Web exhibit a high rate of content churn (Brewington & Cybenko, 2000). Speaking of his survey, Brewington estimated that 20% of Web pages are less than twelve days old, while only 25% are older than one year (Markoff, May 29, 2000). An earlier survey by Douglis, Feldmann & Krishnamurthy (1997) found 16.5% of Web pages to be under constant update. An increasingly large number of Web pages are produced "on the fly" by database manipulations. Sherman (1999) calls this "the invisible Web" and concludes that "this trend is going to make it even harder for search engines to be comprehensive Web indexes." The size of the invisible Web is essentially unknown, but may be vast (Abreu, September 11, 2000). BrightPlanet estimates that the invisible Web is five times the size of the visible Web. "Using Google as a benchmark, that means BrightPlanet would estimate there are about 500 billion pages of information available on the web, and only 1/500 of that information can be reached

via traditional search engines" (Sullivan, August 2, 2000).

As the Web grows in size, timely delivery of content becomes a problem. Fisher (April 17, 2000) describes two strategies used to speed content delivery: caching popular content (the approach used by Inktomi) and using distributed servers (the approach used by Akamai). Many Web pages, therefore, are assemblages of cached and variously distributed material. "When a user in Singapore, say, clicks on a popular page in Yahoo, only the first request goes to Yahoo's server in Palo Alto, Calif.; the balance of the page is then delivered from an Akamai server with the shortest, fastest connection to the person in Singapore" (Fisher, 2000: C1).

Even this cursory review indicates that the Web is a wholly decentralized, rapidly growing, churning phenomenon that springs from many communities, many authors, many languages and points of view. What appears in a Web browser as a static, "document-like" object may have been produced by a combination of dynamic scripts or programs, various database manipulations, with content possibly originating from caching and/or distributed processing. A further complication is that Web browsers, themselves, exhibit idiosyncratic qualities that may alter the appearance of Web pages depending on their abilities to support scripts, Applets, cookies, dynamic HTML, cascading style sheets, extensible markup and so on.

A consideration of the foregoing leads me to conclude that the document paradigm ill suits many Web phenomena, and that the classic example of knowledge representation (i.e., *A librarian giving a subject heading to a book*) may no longer be applicable, or economical, in the Web environment.

Legacy Knowledge Representation Methodologies Applied to the Web

It is a truism, perhaps, that we seldom recognize the radical nature of new technologies and prefer to view them as mere extensions of older, more familiar technologies. This impulse expresses itself in the attempts to catalogue the Web or develop a single subject topical scheme for Web pages.

The NetFirst database is an attempt to catalogue the Web by creating a database of MARC records. To date, volunteer Web surfers have contributed approximately 150,000 MARC records (Greene, June 16, 2000). The CORC project combines the efforts of 489 libraries in 24 countries to build a database of Web pages useful to libraries. "The integration of CORC and WorldCat will create a rich, robust database shared on a global scale, making each library's unique material available to library users worldwide" (Surface, 2000: 33).

CORC has approximately 26,000 records. The present size and rate of growth of the Web described above compared to the small size of these projects underscores the labor-intensive quality of Web cataloguing, and why it is a strategy appropriate for only small pools of relatively static Web content.

Resource discovery on the Web has developed into a major problem with many searches swamped by thousands of false drops. Considerable activity developing metadata schemes has attempted to address this problem. HTML metadata are terms and phrases located in the <HEAD> element of a Web page using the NAME and CONTENT attributes of the <META> element. The ambition of HTML metadata is the addition of subject topical terms and phrases to Web pages, thus emulating the legacy strategy of adding subject topical terms and phrases to cataloguing records.

> Metadata is data about data. The term refers to any data used to aid the identification, description and location of networked electronic resources. Many different metadata formats exist, some quite simple in their description, others quite complex and rich. IFLA Digital Libraries: Metadata Resources

Numerous user communities have attempted to employ metadata schemes to control their particular data, examples being the Nordic Metadata Project, the Arts and Humanities Data Service, and the United States Federal Statistics project. Readers are directed to the IFLANET International Federation of Library Associations and Institutions Web site for a more complete listing. Since 1995, a series of workshops has promoted the Dublin Core Metadata Initiative as the standard metadata tag set. Dillon provides a comprehensive discussion of how Dublin Core Metadata might address the problem of identifying Web resources. He strongly urges a refocus towards the development of a "MARC version of Dublin Core."

To date, cataloguing the Web by deploying meaning-bearing metadata has been meager.

O'Neill, Lavoie and McClain (May 26, 2000) sampled 1, 024 homepages and found only seven using Dublin Core metatags. Lawrence & Giles (1999) reported low metadata use, finding only 0.3% of sites use the Dublin Core metadata standard. At this time no major Web search engine supports topical metadata. (Taylor, April 1, 1999)

> Current metadata usage patterns are a long way from comprehensive document description at the page level. Finally, most metadata usage is still *ad hoc*; with a few exceptions, most sites do not adhere to a well-defined set of metadata elements. (O'Neill, Lavoie and McClain, May 26, 2000)

The idea of a particular user group customizing its data is a powerful one, as is the idea of a controlled set of terms and phrases used to advantage in Web resource discovery. Two false assumptions, however, seem to block the success of current metadata efforts at this time:

- **The False Community Assumption** The legacy methodology of knowledge representation assumes the existence of a class of disinterested information workers to develop and apply subject cataloguing. The decentralized Web lacks such a disinterested class of information workers. Quite the contrary, the Web is composed of millions of individuals who can markup their pages in any manner they wish. Even worse, Web authors, vying for attention to their Web pages,can use meaning-bearing metadata unscrupulously to gain an advantage in site promotion. The Web lacks community norms to prevent this behavior. Search engines avoid meaning-bearing metadata because meaning-discovery algorithms can be spoofed by untrustworthy information (Taylor, April 1, 1999).
- **The False Document Assumption** Current metadata strategies are designed for "high-level document properties" (Lander, 1998). Placing topical terms and phrases in the <HEAD> element of an HTML document assumes that the semantic content of the <HEAD> element will maintain a time-invariant relationship with the semantic content of the <BODY> element. This assumption is reasonable if one conceives of Web pages as merely digitized paper documents. The preceding sketch of Web technology suggests that the legacy metaphor of paper documents and record structures does not fit Web pages very well. While there may always be a residue of static HTML pages, the majority of future Web pages will reflect the efficiencies of database manipulations and extensible markup. For example,Guernsey (July 18, 2000) describes the deconstruction and vending on the Web of "chapters, maps and even paragraphs" that in the legacy information environment would have been indivisible parts of a book.

This review suggests that meaning-bearing metadata would be best employed within a strongly normative community, and in a manner that did not rely on the legacy concept of the document. Extensible markup technologies permit specific communities to set norms as to the structure and semantics of their data, and is furthermore free of any legacy document-like assumptions. In the future, meaning might find a home as a part of extensible markup technologies.

Extensible Information Technologies

HTML mixes content and presentation tags, a design that reflects its original purpose of displaying scientific papers, but makes general data sharing awkward. Separating content from presentation permits data to be gathered without the baggage of presentation tags, and eases the consistent styling of data from different sources. XML heralds the arrival of the "second-generation Web" (Bosak & Bray, May 1999) and "The era of the distributed object" (Cagle, October 26, 1999). Qin (2000) who traces the development of information technology from MARC records to XML.

Extensible Markup Language (XML) 1.0

An XML resource is a file of text strings, a format that facilitates data sharing. The text strings are defined in semantic markup, arbitrary tags that express a particular user's semantics of the data. In Figure 1,

a banana bread recipe is represented with various arbitrary tags of <recipe>, <name> and <ingredient> that might suit a baker's application.

XML elements are modified by attributes, which are string name-value pairs. Figure 2 illustrates three attributes that emulate Dublin Core metadata and one that is user defined :

```
<?xml version="1.0" standalone='yes'?>
<recipe DCCreator="Susan Cheney" DCDate="1990" DCSubject="Cookery(Bread)">
     <name>Banana Bread</name>
          <ingredient>water</ingredient>
          <ingredient>flour</ingredient>
          <ingredient number="4">bananas</ingredient>
</recipe>
```

Figure 2: An XML resource with attributes

In his recipe XML resource, a baker could add many recipes, notes, observations and other types of data.

The revolutionary aspect of XML is the modularization of information. Information presents itself as a self-describing unit that can does not inhibit processing, storing or display. Topical subject qualifiers (e.g., attributes) are placed at the appropriate level of granularity: Recipe qualifiers are placed at the <recipe> level while ingredient qualifiers are placed at the <ingredient> level.

Extensible information technologies antiquate the legacy concept of document:

> As more and more information becomes available in XML format (and as the mechanism for referencing them gets sufficiently defined) then applications become truly transparent to the notion of servers -- a single XML 'document' could conceivably span hundreds or thousands of servers, in such a way that the physical task of locating a document becomes a secondary consideration at best.
> The upshot of this is that the paradigm that we used to think about the Internet, about documents, and about the nature of information changes radically. Agents, XML code blocks that retain their integrity irregardless of their point of origin, roam the Internet as autonomous units in a sea of contextual relationships. (Cagle, October 26, 1999)

This essay questions where meaning may reside when form is gone. The preceding survey of the structural form of XML provides part of the answer. **Meaning resides in the semantic structure of information. Meaning can also reside in the meaning-bearing terms and phrases placed at the appropriate level of granularity that serve to qualify a specific element of information.**

XML Linking Language (XLink) 1.0

XLink is a candidate recommendation as of July 3, 2000 that describes the linking relationships among XML resources. It generalizes the HTML unidirectional links to multidirectional links among two or more resources, or portion of resources. XLinks can be qualified by attributes, thus suggesting another residence of meaning.

Figure 3, an example from the XLink recommendation, illustrates two XLink standard attributes and one user-defined attribute:

```
<my:crossReference
  xmlns:my="http://example.com/"
  my:lastEdited="2000-06-10"
  xmlns:xlink="http://www.w3.org/1999/xlink"
  xlink:type="simple"
  xlink:href-"students.xml">
Current List of Students
</my:crossReference>
```

Figure 3: An XLink resource with both XLink and non-XLink attributes

XLinks facilitate rich links among extensible resources. Linkages themselves carry some semantic information that can be magnified by terms and phrases that provide a semantic context for linking. XLinks provide another locale for meaning. **Meaning can also reside in the qualifications of the relationships among resources.**

Normative Meaning Communities Using Extensible Information Technologies

Content-bearing metadata may most profitably be employed in a strongly normative community that does not rely on the legacy concept of the document. Examples of strongly normative communities are internet-based electronic marketplaces (Bakos, 1998). The ambition of an electronic marketplace is to share information in the most efficient possible manner, ultimately creating "friction-free" marketplaces for goods and services.

The participants in a marketplace for specific goods and services compose a strongly normative community founded on the trust required in selling/buying transactions. Spoofing behavior, tolerated in a random group of Web pages, would be penalized.

Many user communities are developing their own metadata, as opposed to using the Dublin Core set. The appropriate place of these metadata qualifiers is not in the <HEAD> element of an HTML document, but as element attributes in extensible information resources as illustrated in Figure 2. A motivating feature of this employment of metadata is that it is at the appropriate level of granularity, targeting only specific descriptions of goods and services.

Table 1 lists several consortia that promote electronic marketplaces. Consortia provide services such as the registration of XML schemas. A repository of XML schemas provides models for the newcomers and detailed specifications for sharing or searching for data. As an example, OASIS, the non-profit XML interoperability consortium, maintains an XML.ORG Registry

> In the five days since we began accepting registrations, OASIS has had organizations and companies from Australia, Canada, Germany, India, Japan, Korea, Pakistan, the Ukraine and the United States--all wanting to register as submitters. We are working now to validate their submissions and will be soon be inviting users to access the XML.ORG Registry to find schemas for their particular needs. Goldfarb (June 26, 2000)

BizTalk	A Microsoft-backed consortium for the development and distribution of the BizTalk flavor of business-oriented XML schemas
CommerceNet	Defines specifications to facilitate the interoperability of information and integration of content and services across and between vertical markets
FinXML	A consortium supporting the creation and management of the FinXML language for the integration and exchange of digital information in capital markets
Organization for the Advancement of Structured Information Standards	Nonprofit, international consortium of companies and organizations dedicated to accelerating the adoption of product-independent formats based on public standards
RosettaNet	Standardizes the mechanisms used to define the business processes of vertical markets

Table 1: Consortia promoting extensible information technologies

Table 2 gives examples of extensible information initiatives beyond consortia.

Get There	Internet-based B2B travel procurement solutions for corporations, travel suppliers, portals and corporate mobile travelers
Acord Software Directory	Financial services industry
Commerce XML	Commerce resources
Financial Information eXchange	Real-time electronic exchange of securities transactions
American Institute of Certified Public Accountants	XML-based specification for the preparation and exchange of financial reports and data
adXML	An international, open standard organization, which is defining an advertising XML schema for both on-line and off-line media
loanupdate	Collaborative transaction management product for the mortgage industry

Table 2: Example normative communities based on extensible information technologies

Situational Expertise

Knowledge sharing sites on the Web function as forums or brokers for the exchange of expert or everyday wisdom. MindCrossing may be considered a model. It has a stable of subject topical experts who have created "MindStores." A MindStore is a Web site with articles, best practices, case studies, technical specifications and so on. Some of this content is free, some requires payment.

Situation expertise is triggered by visiting a Web information marketplace and searching on a term. A responding tablet on the browser screen alerts the novice user that expertise about this subject is available (for example, see the MindCrossing demonstration)

Context-sensitive situational expertise can orient users to appropriate metadata, concepts and technical vocabulary.

Allexperts	Created in 1998, was the very first large-scale question and answer service
Askme	Provides custom answers to specific questions
Epinions	Offers unbiased advice on over 100,000 products and services
EXP	EXP connects individuals to experts in hundreds of categories

Table 3: Sources of situation expertise

Busch & Reisman suggest that the most successful Web marketplaces are those that develop "deep, industry-specific knowledge or specialized, industry-specific supply-chain capabilities." The integration of knowledge representation and extensible information technologies, combined with situated expertise may facilitate such deep, industry-specific knowledge.

Conclusion

Legacy knowledge representation methods reflect the antiquated paradigm of massive, singular databases of highly structured, identical records. By contrast, extensible information technologies are creating new ways of structuring information and linking information resources.

Extensible information technologies enjoy significant advantages such as the modularization of information, semantic information structures, qualifiers (i.e., content-bearing metadata) placed at the appropriate level of granularity, and semantic relationships among information resources.

As we lose familiar metaphors such as the "document" and address ourselves to the problem of locating meaning in the second-generation Web, the future residences of meaning may be:

- The structure of an information store
- Qualifiers of elements of an information store
- Relationships among information stores
- Expert opinion interpreting the structure, qualifiers and relationships of an information store.

References

- Abreu, E. (September 11, 2000). "Diving into the deep web." *The Industry Standard*, **3**(350, 119.
- Bakos, Y. (1998). "The emerging role of electronic marketplaces on the Internet." *Communications of the ACM*, **41**(8), 35-42
- Bosak, J. and Bray, T. (May 1999). "XML and the second-generation Web." *Scientific American.* [Accessed July 7, 2000]
- Brewington, B. E. and Cybenko, G. (January 29, 2000). "How dynamic is the web?" [Accessed June 28, 2000]
- Busch, J. and Reisman, L. (July 24, 2000). "B-to-B exchanges: know your domain" *The Industry Standard*, **3**(27), 96.
- Cagle, K. (October 26, 1999). "Why XML? A look at XML and how it will change the world." [Accessed June 19, 2000]
- Dillon, M. (2000). "Metadata for web resources: How metadata work on the web." [Accessed September,

2000]
- Douglis, F., Feldmann, A., & Krishnamurthy, B. (1997). "Rate of change and other metrics: A live study of the World Wide Web." [Accessed June, 2000]
- Fisher, L.M. (April 17, 2000). "2 companies take separate paths to speed delivery of Web pages." *New York Times*, p. C1-C4.
- Goldfarb, C. F. (Monday, June 26, 2000). "XML community rallies behind XML.ORG Registry public clearinghouse for XML schemas and vocabularies takes off." *XML Times*, [Accessed June 26, 2000]
- "Google Launches World's Largest Search Engine." (June 26, 2000). [Accessed June 28, 2000]
- Greene, R. (June 16, 2000). "Database load and diacritics and UNICODE." Personal e- mail from Richard Greene, greenr@OCLC.ORG
- Guernsey, L. (July 18, 2000). "Books by the chapter or verse arrive on the Internet this fall." *New York Times*, p. 1-C6.
- Digital Libraries:Metadata Resources IFLANET: International Federation of Library Associations and Institutions. [Accessed July 7, 2000]
- Lander, R. (January 1, 1998). "The search for metadata." [Accessed June 19, 2000]
- Lawrence, S. & Giles, L. (1999). "Accessibility and distribution of information on the web." *Nature*, **400**, 107-109. [Accessed June, 2000]
- Library of Congress "Subject Headings - Principles of Structure and Policies for Application: Contents" [Accessed June, 2000]
- "The MARC 21 Formats: Background and Principles Revised November 1996." [Accessed June, 2000]
- Markoff, J. (May 29, 2000). "As web expands, search engines puff to keep up." **New York Times**, CXLIX, p. C3.
- Nielsen, J. (1995). "Kill the 53-day meme." [Accessed June 28, 2000]
- O'Neill, E.T., Lavoie, B.F., & McClain, P.D. (May 26, 2000). "An analysis of metadata usage on the web." [Accessed July 7, 2000]
- Qin, J. (2000). "Representation and organization of information in the Web Space: From MARC to XML." *Informing Science*, **3**(2).
- Sherman, Chris (June, 1999). "The Invisible Web." [Accessed June 29, 2000]
- Smith, K. W. (1996). "OCLC - Moving toward the next stage of the electronic library." In 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.
- Sullivan, D. (July 5, 2000). "The search engine report", number 44, part 1 of 2. [Accessed July 6, 2000]
- Sullivan, D. (August 2, 2000). "The invisible Web gets deeper." Accessed August 4, 2000.
- Surface, T. (July/August 2000). "CORC: Build locally, share globally." *OCLC Newsletter*, no. 246.
- Svenonius, E. (2000). The Intellectual foundation of information organization. Cambridge, MA: The MIT Press.
- Taylor, C. (April 1, 1999). "An introduction to metadata." [Accessed March 17, 2000]