Information Research, Vol. 8 No. 3, April 2003

Web search: how the Web has changed information retrieval

Terrence A. Brooks
Information School
University of Washington
Seattle, Washington, USA

Abstract

Topical metadata have been used to indicate the subject of Web pages. They have been simultaneously hailed as building blocks of the semantic Web and derogated as spam. At this time major Web browsers avoid harvesting topical metadata. This paper suggests that the significance of the topical metadata controversy depends on the technological appropriateness of adding them to Web pages. This paper surveys Web technology with an eye on assessing the appropriateness of Web pages as hosts for topical metadata. The survey reveals Web pages to be both transient and volatile: poor hosts of topical metadata. The closed Web is considered to be a more supportive environment for the use of topical metadata. The closed Web is built on communities of trust where the structure and meaning of Web pages can be anticipated. The vast majority of Web pages, however, exist in the open Web, an environment that challenges the application of legacy information retrieval concepts and methods.

Introduction

Search is a compelling human activity that has extended from the Library at Alexandria to the World Wide Web. The Web has introduced millions of people to search. The information retrieval (IR) community stands ready (Bates, July 2002) to suggest helpful strategies for finding information on the Web. One classic IR strategy - indexing Web pages with topical metadata - has already been tried, but the results are disappointing. Apparently, relying on Web authors to garnish their Web pages with valid topical metadata runs afoul of human nature:

- Sullivan (2002c) reports that the meta *keywords* tag, an HTML element designed for adding descriptors to Web pages, is regarded as untrustworthy and avoided by all major search engines.
- A FAQ at the Dublin Core site explains that well-known 'all the Web' search engines 'tend to avoid using the information found in meta elements' for fear it is spam (FAQ).

Applying topical metadata to Web pages provokes a controversy that pits partisans who envisage a semantic Web featuring topic maps and ontologies of shared meanings (Berners-Lee, Hendler & Lassila, May 2001) versus detractors who disdain topical metadata as 'metacrap' (Doctorow, August 26, 2001) and warn us of a Web of deception (Mintz, 2002). The significance of the controversy, however, awaits the examination of a more fundamental issue: does it make *technological* sense to add topical metadata to Web pages?

If the Web is a big, distributed document database and Web pages are composed in HTML (i.e., 'the document in my browser goes from <html> down to </html>'), the answer is ‘yes.’ In this case, it makes technological sense for Web authors to add topical metadata to Web pages, just as an indexer might add descriptors to a document in a database. An affirmative answer validates the topical metadata debate. If, however, the Web is not a big document database, but is instead a network of rapidly changing presentations, the answer is ‘no.’ In this view HTML is primarily a presentation technology, and most Web pages are transitory and volatile presentations governed by the whims of viewer taste and the contingencies of viewer technology. A negative answer signals that debating the value of topical metadata is premature until it can be shown that they are technologically appropriate additions to Web pages.

Lurking behind the topical metadata controversy is our unsteady application of the concept of 'document' to Web content and presentation. We inherit our notion of document from vertical-file systems and document databases, technological environments not known for schisms between content and presentation. Viewed from the document-database tradition, indexing Web pages appears to be a simple extension of current practice to a new, digital form of document. Viewed from the HTML tradition, however, indexing Web pages confuses presentation for content. Topical metadata are intended to index information content, not arbitrary or personalized views of content, and the majority of Web pages are arbitrary presentations contingent on Web browsers, security settings, scripts, plug-ins, cookies, style sheets and so on.

Considering the appropriateness of the document metaphor for the Web has fundamental consequences for the application of IR's extensive body of theory and practice. Controversies about topical metadata aside, recognizing the familiar IR notion of 'document' on the Web would suggest that Web searchers are retrieving information, and that we can apply IR concepts and methods to help Web searchers. In this case, the topical metadata controversy gains significance. Realizing that the document metaphor does not map to the Web, however, heralds a paradigm shift. Perhaps Web searchers are not retrieving information, but doing something else. 'Web search' is used in this essay to name the activity of discovering, not retrieving, information on the Web.

IR and the 'document' metaphor

The technological legacy of search

The foundation of search in the last century has been the storage and retrieval of paper based on some form of labeling. Yates (2000) describes vertical filing that made information accessible by using labeled files to hold one or more papers:

> Vertical filing, first presented to the business community at the 1893 Chicago World's Fair (where it won a gold medal), became the accepted solution to the problem of storage and retrieval of paper documents….The techniques and equipment that facilitated storage and retrieval of documents and data, including card and paper files and short- and long-term storage facilities, were key to making information accessible and thus potentially useful to managers. (Yates, 2000: 118 -120)

The application of computer databases to search by mid-20th century extended the vertical file paradigm of storage and retrieval. A computer database is a storage device resembling a vertical file just as a database record is a unit of storage resembling a piece of paper. The more abstract term 'document' addressed any inexactitude in the equivalence of 'database record = piece of paper." Computer databases were seen as storing and retrieving documents, which were considered to be objects carrying information:

- 'Information retrieval is best understood if one remembers that the information being processed consists of documents.' (Salton & McGill, 1988, p. 7)
- 'With the appearance of writing, the document also appeared which we shall define as a material carrier with information fixed on it.' (Frants *et al.*, 1997: 46)
- 'Document: a unit of retrieval. It might be a paragraph, a section, a chapter, a Web page, an article, or a whole book.' (Baeza-Yates & Ribeiro-Neto, 1999: 440)

Digitizing documents greatly boosted the systematic study of IR. Texts could be parsed to identify and evaluate words, thereby perhaps discovering meaning. Facilitating assumptions about the nature of documents and authorial strategies were advanced. For example, Luhn (1959: 160) suggested that 'the frequency of word occurrence in an article furnishes a useful measurement of word significance.' In the following extract Salton and McGill (1988) suggest where subject topical terms are located in documents, and how text can be processed to find these terms:

> The first and most obvious place where appropriate content identifiers might be found is the text of the documents themselves, or the text of document titles and abstracts…. Such a process must start with the identification of all the individual words that constitute the documents…. Following the identification of the words occurring in the document texts, or abstracts, the high-frequency function words need to be eliminated… It is useful first to remove word suffixes (and possibly also prefixes), thereby reducing the original words to word stem form. (Salton & McGill, 1988: 59, 71).

The document-database search technology sketched above maps easily to the Web and suggests that searching on the Web is an extension of IR:

- Vast numbers of documents are available on the Web (e.g., 'the Web is a big database.')
- Viewing the source of a Web presentation reveals a structured document (e.g.: 'the document goes from <html> down to </html>.')
- Google seems to index Web pages (e.g., 'Google is a big index made up of words found in Web pages.')

The legacy social context of search

We inherit, as well, an elaborate social context of search that has been applied to the Web. Librarianship was the source of powerful social conventions of search even before the introduction of the technology of vertical files. For example, Charles A. Cutter suggested rules for listing bibliographic items in library catalogues as early as 1876. Bibliographic standardization, expressed in the Anglo-American Cataloguing Code, was a powerful idea that promoted the view that the world could cooperate in describing bibliographic objects. An equally impressive international uniformity was created by the wide acceptance of classification schemes, such as the Dewey Decimal Classification (DDC):

> Other influences are equally enduring but more invisible, and some are especially powerful because they have come to be accepted as 'natural.' For example, the perspectives Dewey cemented into his hierarchical classification system have helped create in the minds of millions of people throughout the world who have DDC-arranged collections a perception of knowledge organization that had by the beginning of the twentieth century evolved a powerful momentum. (Wiegand, 1996: 371)

The application of computer databases by mid-20th century spurred many information communities to establish or promote social conventions for their information. For example, the Education Resources Information Center (ERIC), 'the world’s largest source of education information' (Houston, 2001: xiv), represents a community effort to structure and index the literature of education. At the height of the database era in the late 1980s, vendors such as the Dialog Corporation offered access to hundreds of databases like ERIC, each presenting one or more literatures structured and indexed. This social cooperation and technological conformity fostered the impression that, at least in regards to certain subject areas, the experts had their information under control.

The social context of document-database search sketched above maps easily to the Web and suggests a benign, socially cooperative information environment:

- Web authors will add topical metadata to their Web pages (e.g., 'I index my Web pages with keywords and Dublin Core metadata so people will find them on the Web.')
- Everyone will use topical metadata (e.g., 'The semantic Web will be constructed by millions of Web authors indexing their Web pages.')
- Web crawlers, like Google, will harvest topical metadata (e.g., 'Google has indexed my topical metadata and now my Web pages are available for retrieval.')

We are now just learning that the Web has a different social dynamic. The Web is not a benign, socially cooperative environment, but an aggressive, competitive arena where authors seek to promote their Web content, even by abusing topical metadata. As a result, Web crawlers must act in self defense and regard all keywords and topical metadata as spam.

Debating whether topical metadata are spam or an essential step towards the construction of the semantic Web assumes that they are technologically appropriate additions to Web pages. To what extent are Web pages analogues of the legacy IR document-container of information?

The Web and the 'document' metaphor

A Web page is a 'snapshot'

Documents added to the ERIC database thirty years ago are still retrievable. There is every expectation that they can be retrieved next year. This expectation provides a rough definition of what it means to retrieve information – finding the same document time and again. The metaphor used in the working draft on the *Architectural Principles of the Web* (Jacobs, August 30, 2002), however, does not suggest retrieving the same thing time and again.

Interacting with a Web resource gives one a snapshot:

> There may be several ways to interact with a resource. One of the most important operations for the Web is to retrieve a representation of a resource (such as with HTTP GET), which means to retrieve a snapshot of a state of the resource. (Jacobs, August 30, 2002, section 2.2.2)

Web resources are characterized as evolving, not static, resources. They are more like loose-leaf binder services than time-invariant database records:

> An integrating resource is a bibliographic resource that is added to or changed by means of updates that do not remain discrete and are integrated into the whole. Examples of integrating resources include updating loose-leafs and updating Web sites. (Task group on implementation of integrating resources, 2001)

If Web pages are snapshots then the critical question is rate of update. Some ERIC records are thirty years old; the oldest HTML pages date from about ten years ago, but most Web content is much more volatile:

- Brewington and Cybenko (2000) observed that half of all Web pages are no more than 100 days old, while only abut 25% are older than one year.
- Cho and Garcia-Molina (2000) found 40% of Web pages in the *.com* domain change everyday. The half-life of Web pages in the *.gov* and *.edu* is four months.
- Koehler (1999) found the half-life of Web content is two years.
- Spinellis (2003) found the half-life of URLs is four years.
- Markwell and Brooks (April 15, 2002) found the half-life of science education URLs to be fifty-five months.
- Cockburn and McKenzie (2001) found that the half-life of bookmarks to be two months.

Content churn and rapid birth and death cycles distinguish Web pages from the legacy IR document-container of information. Philosophers can address the issue of repeated refreshing of the 'same' Web page that presents 'different' content each time, as to whether this is the 'same' Web page or 'different' Web pages. Whatever grist falls from the philosophical mill, it is clear that Salton and McGill didn’t consider database documents to be snapshots.

Web pages are cultural artifacts

Web content is only available through the mediation of a presentation device, such as a Web browser. Complicating Web presentation are security settings, different computer monitors, safe and unsafe Web colors, plug-ins, cookies, scripts, and so on. In fact, Web authors expend enormous amounts of time and energy engineering a consistent presentation across platforms.

> The representations of a resource may vary as a function of factors including time, the identity of the agent accessing the resource, data submitted to the resource when interacting with it, and changes external to the resource.' (Jacobs, August 30, 2002, section 2.2.5)

Figure 1 illustrates the process of converting HTML to a browser display for the Mozilla layout engine (Waterson, June 10, 2002).

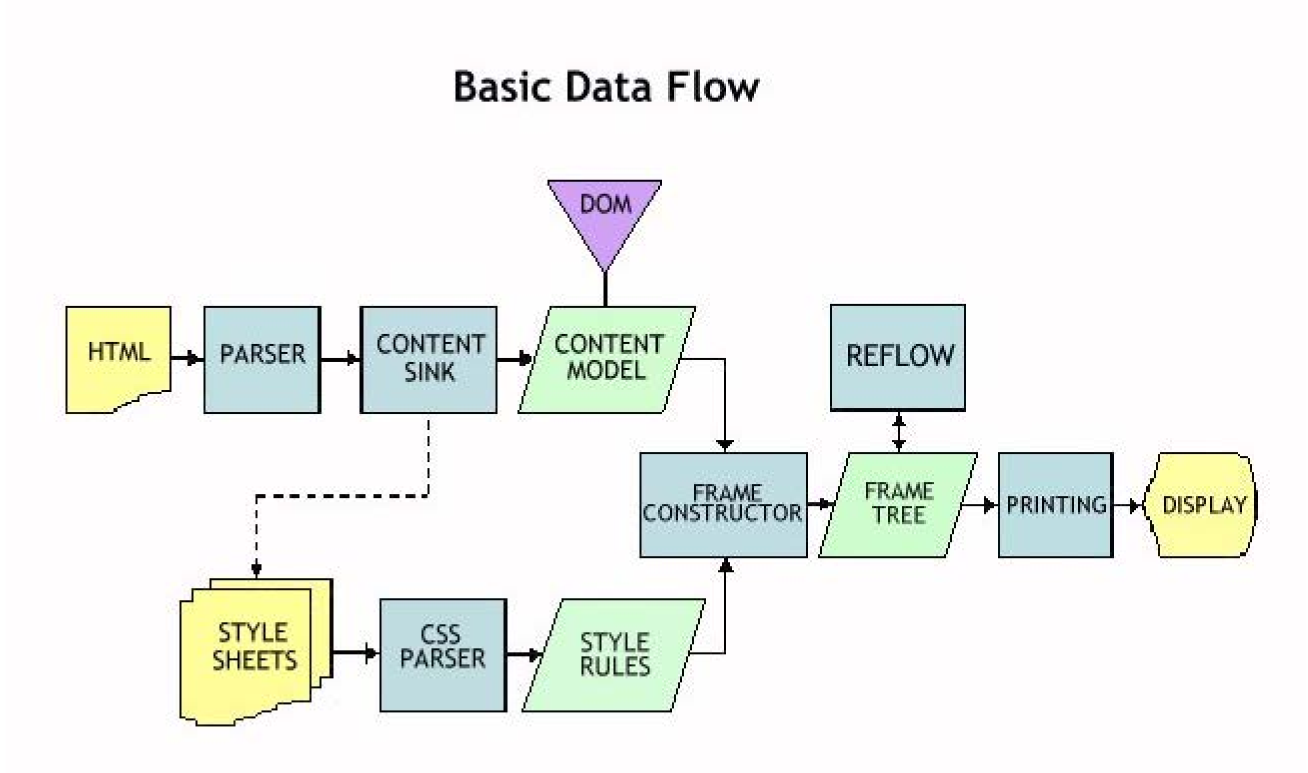

Figure 1: Basic data flow in Mozilla layout engine (Waterson, June 10, 2002)

This diagram illustrates that HTML code is parsed and deconstructed into a hierarchical content model. Style sheets that reference content elements are also parsed. A frame constructor mixes content with style rules into a hierarchy of content frames. Nested content frames are painted to create the presentation in a Web browser. Figure 1 implies that different HTML parsing rules, style sheet applications, frame construction algorithms, and so on, would produce a different presentation.

Figure 1 also implies that assembling Web content to look like a printed document is not a technical necessity, but a cultural convention. If your Web browser presents you with something that looks like a printed page, it is because the engineers of Web browsers are obeying the cultural expectations of the majority of their users; that is, information should resemble the familiar printed page. The mutability of Web presentation is not deplored, but actually trumpeted as an advantage in delivering customized appearance. In short, Web content can be made to look like your *favorite printed page*:

> Cookies serve to give Web browsers a ‘memory’, so that they can use data that were input on one page in another, or so they can recall user preferences or other state variables when they user leaves a page and returns. (Flanagan, 1997, p.231)

Finding the Web page in your browser located between <HTML> and </HTML> tags reflects how your browser constructed the byte stream from the source server machine, but says nothing about how the content was structured on the source server machine. The source content could be distributed among a number of databases, XML documents, scripts, files and so on. XSLT style sheets can assemble Web site 'skins' from databases and XML sources with equal ease (Pierson, March 2003)

During the early years of the Web, most Web pages were constructed in HTML and many handcrafted Web pages are still written this way. Efficiencies of scale, however, have forced large producers of Web content to automate Web page production:

- Turau (1999) speculated that 75% of Web pages are generated from databases.
- Bergman (2001) describes the 'deep' Web as 400 to 500 times larger than the 'surface' Web. The deep Web is composed of database generated pages.

Web pages are presentation contingencies and server programming artifacts. This schism between content and presentation distinguishes them from the legacy IR notion of document-container of information. The document in your Web browser may look like a document, but probably has no documentary origin at all.

Google is not a Web index

An index helps searchers find information in a database. It is generally true that success in finding information in a database is directly proportional to knowledge about how the database documents were indexed. Database vendors such as the Dialog Corporation are famous for running classes helping searchers understand how information is indexed. Google is a popular search tool for Web content, twice voted most outstanding search engine by the readers of Search Engine Watch. In August 2002, about 28% of Web search was done with Google (Sullivan, 2002b). 'Google gets 150 million queries a day from more than 100 countries' (Harmon, February 17, 2003). Google is famous for presenting results according to page rank, but nobody knows how Google's parsing algorithm works.

Sullivan (2002a) surmises that Google uses over 100 factors to parse Web content, which still includes 'traditional on-the-page factors.' (The algorithm of Salton and McGill focuses on these factors.) If Google were to expose its parsing algorithm, it would be immediately exploited by Web authors seeking to gain advantage and visibility for their Web content. Google's economic viability depends maintaining this secret: a corporate strategy strikingly different from legacy database vendors like the Dialog Corporation. Google warns Web authors who would attempt to ferret out and exploit their parsing algorithm:

> We will not comment on the individual reasons a page was removed and we do not offer an exhaustive list of practices that can cause removal. However, certain actions such as cloaking, writing text that can be seen by search engines but not by users, or setting up pages/links with the sole purpose of fooling search engines may result in permanent removal from our index. (http://www.google.com/Webmasters/2.html).... Google's complex, automated methods make human tampering with our results extremely difficult (http://www.google.com/technology/index.html).

Google does not attempt to cover the entire Web. It systematically excludes Web sites with doorway pages or splash screens, frames and pages generated 'on-the-fly' by scripts and databases. Jesdanun (October 25, 2002) reports content removed from Google to satisfy national prohibitions. Many Web pages also include objects Google finds opaque such as image files and applets. Increasing numbers of Web presentations have no text at all: 'Graphic design can be content where users experience a Web-site with little or no "text" *per se*'. (Vartanian, 2001).

Legacy indexing algorithms were open for inspection. Adding topical metadata to Web pages in the hope that Google will harvest them is betting on an unknown indexing strategy. Google will no tell you what it did with your topical metadata because being a black box is a corporate survival strategy.

Conclusion

The preceding survey of Web technology indicates that Web pages make poor hosts for topical metadata. This is not an evaluative judgment about topical metadata themselves, but merely an observation that they are misapplied to a technology characterized by churning content in arbitrary presentations parsed by unknowable algorithms. The cost and effort of adding topical metadata to an information structure is only recouped if that information structure persists in time with a predictable structure, identity and contents. An example of such a structure is the legacy IR document-container of information. Topical metadata await their more appropriate application on the Web in environments where the technical and social factors supporting the IR document-container of information can be re-created. This can be done by 'closing' the Web.

Closing the Web to do information retrieval

The legacy technical and social environments supporting IR in document databases are sketched above. It is possible to re-create this environment on the Web behind passwords in venues such as intranets, enterprise computing, and digital libraries. These applications are driven by social groups that can reach agreements on information structure and topical metadata. For example, a social group can arbitrarily decide to construct and present its information in HTML or any other presentation technology. It can also decide to use the meta keywords tag or Dublin Core metadata with its choice of thesaurus of indexing terms and phrases.

Social agreements take precedence over technology in closed Webs where the Web is reduced to a communications venue. Predictability in structure and meaning is the fundamental facilitator permitting in-house Web crawlers to harvest topical metadata for the retrieval benefit of the local community. In a closed Web, one can build a legacy database and do IR.

If our terms of reference are a closed Web, then the topical metadata controversy recognizes topical metadata as important elements of a semantic (i.e., 'closed') Web.

Web search on the 'open' Web

Now and in the future, Google and similar tools will scan billions of Web presentations on the open Web. Everyday searchers will use Google to find information, and many will characterize their activity as retrieving information, despite disappearing Web pages, rotten links and Web content that changes on each viewing.

The open Web is a network where the cost of entry is merely access to a server machine. There are no social conventions about who can author a Web presentation or what can be presented. It is an unconstrained environment where initiatives requiring Web authors to add indexing terms and phrases or to structure their Web pages a certain way are doomed to failure, or will be exploited by the unscrupulous. In the open Web there is no guarantee that Web presentations will remain or that servers will continue to function.

Many will use the Web to 'retrieve information,' but they are engaged in Web search, a process of constant discovery, not retrieval. The only way to preserve a Web presentation is to cache it, which is to take a *snapshot of a snapshot* and thereby create a new static representation of a continuously evolving process.

The open Web challenges us to ransack our IR legacy of concepts and methods to find any that can be applied. But it is possible that the open Web is so novel a technological platform that we will be forced to recognize that our IR legacy of concepts and methods has been historicized to the modern database era of the late 20th century.

References:

- Baeza-Yates, R. & Ribeiro-Neto, B. (1999) *Modern information retrieval*. Reading, MA: Addison-Wesley.
- Bates, M.J. (2002). After the dot-bomb: getting Web information retrieval right this time. *First Monday*, **7** (7) http://firstmonday.org/issues/issue7_7/bates/index.html (20 March 2003)
- Bergman, M. K. (2001) The deep Web: surfacing hidden information *The Journal of Electronic Publishing*, **7** (1) http://www.press.umich.edu/jep/07-01/bergman.html (18 January 2003)
- Berners-Lee, T., Hendler, J. & Lassila, O. (2001) The semantic Web. *Scientific American*, **284**(5), 34-+ http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2 (20 March 2003)
- Brewington, B.E. & Gybenko, G. (2000) How dynamic is the Web? In: [Proceedings of the] Ninth International World Wide Web Conference, Amsterdam May 15-19, 2000. WWW9.org. http://www9.org/w9cdrom/264/264.html (20 March 2003)
- Cho, J. & Garcia-Molina, H. (2000) The evolution of the Web and implications for an incremental crawler. In: Amr El Abbadi, Michael L. Brodie, Sharma Chakravarthy, Umeshwar Dayal, Nabil Kamel, Gunter Schlageter, and Kyu-Young Whang, editors. *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt*. San Francisco, CA: Morgan Kaufmann. http://rose.cs.ucla.edu/~cho/papers/cho-evol.pdf (20 March 2002)
- Cockburn, A. & McKenzie, B. (2001) 'What do Web users do? An empirical analysis of Web use' *International Journal of Human-Computer Studies*, **54**(6), 903-922
- Doctorow, C. (2001) Metacrap: putting the torch to seven straw-men of the meta-utopia. [Personal Web site] http://www.well.com/~doctorow/metacrap.htm (20 March 2003)
- FAQ: What search-engines support the Dublin Core Metadata Element Set? Dublin Core Metadata Initiative. http://dublincore.org/resources/faq/#whatsearchenginessupport (20 March 2003)
- Flanagan, D. 1997. *JavaScript: the definitive guide*. Sebastopol, CA: O'Reilly.
- Frants, V.I., Shapiro, J. & Voiskunskii, V.G. (1997) *Automated informaton retrieval: theory and methods.* San Diego, CA: Academic Press.
- Harmon, A. (2003) Google deal ties company to Weblogs. *The New York Times* Monday, February 17, C3.
- Houston, J. E. (2001) *Thesaurus of ERIC descriptors*, 14th edition. Westport, CT: Oryx Press.
- Jacobs, I. (2002). *Architectural principles of the World Wide Web, W3C working draft.*. World Wide Web Consortium (W3C.org) http://www.w3.org/TR/2002/WD-Webarch-20020830/ (20 March 2003)
- Jesdanun, A. (2002) Report: sites missing from Google. News.com.au

http://www.news.com.au/common/printpage/0,6093,5356761,00.html (20 March 2003)

- Koehler, W. (1999) Digital libraries and the World Wide Web sites and page persistence. *Information Research*, **4**, (4) http://informationr.net/ir/4-4/paper60.html (23 March 2003)
- Luhn, H. P. (1959) The automatic creation of literature abstracts. *IBM Journal of Research and Development* **2**(2), 159-165.
- Markwell, J. & Brooks, D.W. (2002) Broken links: just how rapidly do science education hyperlinks go extinct? Lincoln, NE: University of Nebraska-Lincoln. Department of Biochemistry. http://www-class.unl.edu/biochem/url/broken_links.html (20 March 2003)
- Mintz, A. (2002) Web of deception: misinformation on the Internet. Medford, NJ: Information Today.
- Pierson, H. (2003) Site skinning: rich XML classes let users personalize their visual experience on your ASP.NET site. *MSDN Magazine*, March, 87-92. http://msdn.microsoft.com/msdnmag/issues/03/03/SiteSkinning/default.aspx (20 March 2003)
- Salton, G. & McGill, M.J. (1988) *Introduction to modern information retrieval*. New York, NY: McGraw-Hill.
- Spinellis, D. (2003) The decay and failures of Web references. *Communications of the ACM*, **46**(1), 71-77
- Sullivan, D. (2002a) Google: Can the Marsha Brady of search stay sweet? Search Engine Watch. http://searchenginewatch.com/sereport/02/09-google.html (20 March 2003)
- Sullivan, D. (2002b) Nielsen//NetRatings search engine ratings. Search Engine Watch. http://www.searchenginewatch.com/reports/netratings.html (20 March 2003)
- Sullivan, D. (2002c) Death of a meta tag Search Engine Watch http://searchenginewatch.com/sereport/02/10-meta.html (20 March 2003)
- Task Group on Implementation of Integrating Resources. (2001) *Interim report.* Washington, DC: Library of Congress. http://www.loc.gov/catdir/pcc/tgintegrpt.html (20 March 2003)
- Turau, V. (1999) Making legacy data accessible for XMl applications. [Personal Web site] http://www.informatik.fh-wiesbaden.de/~turau/ps/legacy.pdf (20 March 2003)
- Vartanian, I. (2001) *Now loading....* Coret Madera, CA: Gingko Press
- Waterson, C. (2002) Introduction to layout in Mozilla. Mozilla.org http://www.mozilla.org/newlayout/doc/gecko-overview.htm (20 March 2003)
- Wiegand, W.A. (1996) *Irrepressible reformer, A biography of Melvil Dewey*. Chicago, IL: American Library Association.
- Yates, J. (2000) Business use of information and technology during the industrial age, in *A Nation transformed by information: how information has shaped the United States from Colonial Times to the Present*, edited by A.D. Chandler & J.W. Cortada. pp. 107-136. New York, NY: Oxford University Press.

Find other papers on this subject.

(Web OR "World Wide Web") searching "information retrieval"

Scholar Search

(Web OR "World Wide Web") searching "information retrieval"

Google Search

How to cite this paper:

Brooks, Terrence A. (2003) "Web Search: how the Web has changed information retrieval" *Information Research*, **8**(3) paper no. 154 [Available at http://InformationR.net/ir/8-3/paper154.html]

Last updated: 20 March 2003

Check for citations, using Google Scholar