Information Research, Vol. 4 No. 4, July 1999

Digital libraries and World Wide Web sites and page persistence.

Wallace Koehler
School of Library and Information Studies
University of Oklahoma
Norman, OK, USA

Abstract

Web pages and Web sites, some argue, can either be collected as elements of digital or hybrid libraries, or, as others would have it, the WWW is itself a library. We begin with the assumption that Web pages and Web sites can be collected and categorized. The paper explores the proposition that the WWW constitutes a library. We conclude that the Web is not a digital library. However, its component parts can be aggregated and included as parts of digital library collections. These, in turn, can be incorporated into "hybrid libraries." These are libraries with both traditional and digital collections. Material on the Web can be organized and managed. Native documents can be collected *in situ*, disseminated, distributed, catalogueed, indexed, controlled, in traditional library fashion. The Web therefore is not a library, but material for library collections is selected from the Web. That said, the Web and its component parts are dynamic. Web documents undergo two kinds of change. The first type, the type addressed in this paper, is "persistence" or the existence or disappearance of Web pages and sites, or in a word the lifecycle of Web documents. "Intermittence" is a variant of persistence, and is defined as the disappearance but reappearance of Web documents. At any given time, about five percent of Web pages are intermittent, which is to say they are gone but will return. Over time a Web collection erodes. Based on a 120-week longitudinal study of a sample of Web documents, it appears that the half-life of a Web page is somewhat less than two years and the half-life of a Web site is somewhat more than two years. That is to say, an unweeded Web document collection created two years ago would contain the same number of URLs, but only half of those URLs point to content. The second type of change Web documents experience is change in Web page or Web site content. Again based on the Web document samples, very nearly all Web pages and sites undergo some form of content within the period of a year. Some change content very rapidly while others do so infrequently (Koehler, 1999a). This paper examines how Web documents can be efficiently and effectively incorporated into library collections. This paper focuses on Web document lifecycles: persistence, attrition, and intermittence. While the frequency of content change has been reported (Koehler, 1999a), the degree to which those changes effect meaning and therefore the integrity of bibliographic representation is yet not fully understood. The dynamics of change sets Web libraries apart from the traditional library as well as many digital libraries. This paper seeks then to further our understanding of the Web page and Web site lifecycle. These patterns challenge the integrity and the usefulness of libraries with Web content. However, if these dynamics are understood, they can be controlled for or managed.

Introduction

The World Wide Web offers a challenge unlike traditional media to those who seek to manage it and to categorize it (Johnston, 1998). The Web is different. I have argued elsewhere that prior to the advent of the Internet, information

was managed in essentially two ways. The first is ephemeral, centrally owned and controlled, and unrecorded. Oral traditions, the spoken word, and even live unrecorded broadcasts are examples. The second is the written or recorded tradition. These materials are more permanent not only because they are stored in some format, but also because their ownership tends to be diffuse. Web content lies somewhere between these two traditional information storage and dissemination strategies. Web content may be centrally owned but universally accessible. At the same time, edited Web material replaces its antecedent, usually leaving no trace of the previous document/edition. Once a Web document is permanently removed from the WWW it ceases to exist (Koehler, 1997). Our understanding of Web dynamics is further complicated by the proliferation of digital libraries of varying quality, complexity, and scope (Lesk, 1996) and a definitional confusion between digital libraries and the Web.

We begin with the assumption that Web pages and Web sites can be collected and categorized. The paper explores the proposition that the Web constitutes a library. We conclude that the Web is not a digital library. However, its component parts can be aggregated and included as parts of digital library collections. These, in turn, can be incorporated into "hybrid libraries." These are libraries with both traditional and digital collections (Pinfield, *et al.*, 1998). Material on the Web can be organized and managed. Native documents can be collected *in situ*, disseminated, distributed, catalogueed, indexed, controlled, in traditional library fashion. The Web therefore is not a library but library material can be selected from the Web.

This paper seeks then to further our understanding of the dynamics of Web page and Web site lifecycles. These patterns challenge the integrity and the usefulness of libraries with Web content. This paper focuses on individual Web page and site attrition rather than the Web as a whole. Web pages and Web sites can be collected and categorized. They can be aggregated and included as parts of library collections. These libraries can take one of two forms. Web libraries can consist of native documents linked to the library, which provides access to the document but do not exercise control or ownership over the document. Web libraries can also "collect" Web documents through caching or archiving, thus creating a local electronic record on an historical Web document.

The Web as Library

Is the Web a library? Is it a digital library? Digital libraries, including Web-based collections, are redefining both the role of electronic information storage and retrieval as well as the role of traditional libraries. As Ching-chih Chen (1998) has argued, no cohesive or comprehensive theory of digital libraries has yet to be fully developed. Digital libraries include collections of books, journal articles, graphics, newspapers, and other material in digital format; in sum, collections of digitized content. To be a library, that collection should be organized according to some standard (Chen, 1998).

There are many examples of digital libraries (Digital Library Information and Resources 1999). These include Project Gutenberg, the US Library of Congress National Digital Library Program, online journal collections, and so on. These libraries, like traditional libraries, vary in their collections and their complexity (Bailey, 1999). The Digital Library Federation offers one definition of digital libraries:

> Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities.

Because we sometimes try to treat information on the Web similarly to the way we organize information in a library, does that necessarily mean that the Web is a library? On one level, the Web can be seen as an extension of and as a tool for libraries. If the Web is a tool, it is a complex tool serving to modify the information management and retrieval functions of both librarians and libraries (Arnold, 1994; Prescott, n.d.; Abramson 1998). The Web can, for example, be employed as a transfer and distribution mechanism for the digital and digitized collections of libraries (Lesk 1996). The Web as communications medium has become an interface, if not the major interface to any number of database vendors including Dialog, Lexis/Nexis, Thomson and Thomson, and FirstSearch. Many academic and public libraries provide Web access to their OPACs. Document delivery can be accomplished over the Web, and may include both native Web documents as well as digitized print material (as the search engine and document delivery service Northern Light illustrates). Some, like the US Library of Congress, are beginning to provide Web access to their digitized collections.

The Web is a complex construct for both information transfer and as a repository of information content. Because

the Web represents an information source, is it necessarily a library of those information resources? There are a number of projects, OCLC's NetFirst (Koehler & Mincey, 1996) and the Dublin Core Metadata Initiative (1999), among them, that treat Web content as "library content" by either catalogueing native documents or by providing automated metadata templates. There many other initiatives to bring some kind of order to the information chaos that the Web represents (Oder, 1998).

Web users can organize Web documents into document "clusters" that can be managed, searched, and accessed. These clusters or libraries have taken many forms. The vast majority of these attempt an *in situ* organization and categorization of Web content. Perhaps the most common are the "jumppages," or Web page collections of hypertext links that point to other Web pages or Web sites with a defined content. NetFirst, a part of OCLC's FirstSearch family of databases, employs a MARC format, catalogues Web pages, abstracts and indexes them, and applies a Dewey Decimal Classification for the document. Yahoo! and other Web-based directories index subsets of the Web. The robot indexed search engines, like Alta Vista, Excite, HotBot, Lycos, WebCrawler, and others have created indexes of fulltext and/or metatag keywords or other elements to provide search and retrieval services. There are also a growing number of limited area search engines (LASE) that are restricted to development of indexes for Web documents with specific content. These include *Argos, the Limited Area Search of the Ancient and Medieval Internet* LASE, the *Ananzi* LASE providing access to material from or about South Africa, or the directory of LASEs provided by *Internet Sleuth*.

Is the Web something more than a complex access interface for digital libraries? There are those who argue that the Web is more than a library door or catalogue, it is, in fact, a library (e.g., Travica & Olson, 1998). In Reno v ACLU (1997), the US Supreme Court suggested: "The Web is thus comparable, from the readers' viewpoint, to both a vast library including millions of readily available and indexed publications and a sprawling mall offering goods and services."

It might also represent a precursor to ideas developed by H.G. Wells in his collection of essays published as *World Brain*in 1938. Michael Lesk (1997) suggests that digital libraries share three common traits: (1) they can all be searched, (2) they can all be accessed from anywhere, and (3) they can all be copied using electronic means without error. The Web meets that three-trait test. It can be searched, accessed, and copied.

The Web differs from the more "traditional" information media in several substantial ways. The Web as a whole may not be organized to the degree that Chen might want it to meet the standard of a library, but neither is the entire corpus of print or electronic media so organized either. Part of the problem lies in part in our definition of the Web: sometimes process and sometimes content.

José-Marie Griffiths (1998) offers a four-point argument against the Web as library:

- despite appearances to the contrary, the Web does not offer access to all information;
- the Web lacks authority and quality control;
- the Web is inadequately catalogueed;
- Web search interfaces and other tools are ineffective and simplistic.

The Web lacks authority and quality control, but so do many "print" publications. Much of the Web is reminiscent of the political and religious pamphleteering so prevalent in Europe and North America over the past 200 years. The Web can resemble that archetypal forgery in print *The Protocols of the Elders of Zion* raised to some major order of magnitude. It can be an intentional or unintentional vehicle for poor or wrong information. It is a conduit for pornography. It can also provide effective and timely access to useful and accurate information.

One difficulty with the Web is that it is today far more difficult for third parties to assess quality and authority using methods developed for more traditional publications. T. Matthew Ciolek (1996) describes six approaches to bring quality control to the Web. These are the programming, procedural, structuring, bibliographical, evaluative or indexed, and organizational approaches. Variations and additions to these approaches have been offered as well. Bibliometric or "Webometric" (a term coined by Almind & Ingwersen, 1997) approaches may provide additional quality control. These include citation analysis, link following (Chu, 1997; Baron, *et al.*, 1996) and word counts (Almind & Ingwersen 1997). Others have adopted bibliometric approaches by following hypertext link trails to "authoritative" or substantive sites (Khan & Locatis, 1998). Recent work suggests that following link trails to clusters of clusters (Kaiser, 1998) can identify authoritative Web sites.

The dean of Web quality evaluation, Hope Tilman (1998) suggests a more traditional library approach to quality and authority. She describes time tested criteria, including timeliness, authority, reviews and ratings, and ease of use. She also suggests that format and the "stability of information" are important new criteria. By applying both qualitative and quantitative rigor, progress is being made on quality and authority issues (Auer, 1998 for a bibliography on resource evaluation). There is no doubt that the indexes of the major search engines and directories are inadequate (Tomaivolo & Packer, 1996; Brake, 1997; Lawrence & Giles, 1998). The fact is we do not even know how big the Web is much less how to define "big." The Web contains many documents, pages, sites, and objects. Sizing the Web has proven to be a complex and uncertain enterprise. One Web snapshot (Bray, 1996) put it at approximately 50 million pages in November 1995. Koehler (1998) suggests a count of some 600 million Web pages for late 1996. Lawrence and Giles (1998) offer a 1997 estimate of some 300 million. It may be that the "actual number" may never be countable for a variety of technical and definitional reasons. In addition, it may be that a reasonable approximation will suffice.

There is no agreement as whether the Web can be effectively catalogueed or not (Oder, 1998) or what to catalogue. In an eloquent defense of Web catalogueing, Jul, Childress, and Miller (1997) describe and dispel what are for them the three most persuasive arguments against the practice. These are: (1) Web content is trash, (2) Web content is too ephemeral, and (3) catalogueing technologies were designed for print and are not applicable to the Web. Jul, *et al.*, (1997) are correct. Some Web resources are trash as is some print. But even trash (a highly subjective subject) needs at times to be classified or catalogueed. Their other two observations are more telling. Others recognize that Web document changes are important to their bibliographic control (Hsieh-Yee, 1996) and still others have found it necessary to modify practice in response to those changes (McDonnell, *et al.*, 1999).

Much attention has been paid to Web size and growth, as defined by server and host counts (NetWizards, NetCraft), packet traffic (MIDS) and other indicators. While we recognize both functions, we tend to concentrate more attention on the informatics of the transfer function rather than on the growth and change of the information content of the medium. That said, we are reasonably certain of at least two things: the Web is a "big place" and that it is ever-changing and ever-growing. There have been a number of approaches and proposals to categorize, index, and catalogue the Web (Dillon, *et al.*, 1993; Caplan, 1994; Forrester, 1995; Pattie & Cox, 1996; Olson, 1997; Ellis, *et al.*, 1998; McDonnell, *et al*., 1999; Hsieh-Yee, 1996). One of the more important is Dublin Core, a major and comprehensive effort to describe Web resources (Desai, 1997). Dublin Core (Dublin Core Metadata, 1997) is perhaps the most important to date of the schema developed to capture WWW metadata. The Dublin Core elements list and semantic headers represent a loosely constructed thesaurus in which index terms are author supplied. Dublin Core has been widely adopted and adapted to manage a wide variety of Internet resources in an equally diverse set of geographic and subject environments (Shvartsman, 1998; Nordic, 1998; UKOLN, 1998).

McDonnell, *et al.* (1999) describe one effort to develop a Web catalogue of area studies material, management of the ephemeral character of the collections, and their efforts to apply *AACR2* standards to create records using the MARC template. Because of the ephemeral nature of Web material and the relative durability of Web sites and Web pages located on directory structures closest to the server address, they recommend that individual catalogue entries be general and describe overall content vectors rather than the content of specific Web pages. In a similar vein, Tennant (1998) recommends "digital bibliography" over digital catalogueing where digital bibliography extracts the essence in very general terms of an entire work while digital catalogueing is far more detailed, captures specific metadata, and therefore far more subject to be influenced by change. Woodruff, *et al.* (1996) have categorized documents by: document size, tag/size ratio, tag usage, attribute usage, browser-specific extension usage, port usage, protocols used in child URLs, file types used in child URLs, number of in-links, readability, and syntax errors. In their conclusions they note that the WWW changes rapidly and that therefore document properties often undergo significant changes as well. Koehler (1999b) suggests that URL characteristics (top-level through server-level domain names, ports, file structures), Web site and Web page metrics (size, object counts, object proportions), as well as change characteristics captured over time can be used to categorize Web pages and sites.

Let us concede that the Web is not and is not meant to be a library any more than the authoring, publication, production, distribution, and marketing of "traditional" media constitute libraries. Libraries are collections of these information products, organized in some useful fashion, and made available to the appropriate client or patron pool. Will then the Web become a library if and when it meets Griffiths' tests? Perhaps not or perhaps it does not matter, for libraries can be libraries if their collections either are not universally comprehensive or if their collections do not address all points of view concerning a specific issue. In fact, no library can ever fully meet the "access to all information" test. It can also be argued that in time the Web may more likely provide access to all points of view than print because of the very ease of publication to the Web. Like other information streams, the Web can be

organized, collected, and managed. Libraries, necessarily digital libraries, can be and are constructed from Web resources.

The Web as a River

The Internet is a moving target in two ways. First, there is a constant stream of technical change and computing upgrades and improvements. New capabilities, concepts and terminology challenge the practitioner, but these are manageable (Olson, 1997). The other change is, from the perspective of managing Web content, far more profound. Web pages and Web sites undergo constant modification. These changes take two general forms. The first is existence. Both Web pages and Web sites exhibit varied longevity patterns (Koehler, 1999a). The second is content modification. Koehler also found, for example, that over a period of a year, more than 99 percent of Web pages still extant undergo some degree of change but with varying degrees of frequency.

Efforts to organize, categorize, index, or catalogue the Web all contain one inherent and fatal flaw. Before one can successfully organize Web content, one must recognize and compensate for the dynamic nature of the Web. Like Heraclitus' River, the Web is a constantly flowing and changing river of information. That river differs from its more "traditional" counterparts in some very important ways. Just as the Web is much easier to publish to than the traditional media is; it is also much easier to "unpublish to," and to edit and modify than those same "traditional" media (Koehler, 1999a) are. Not only (first) does the Web undergo frequent content change, (second) it also undergoes frequent death or "comatoseness." These phenomena can be measured, documented, and reported (Koehler, 1999a).

Research problem

Nothing is forever. Some things last longer than others do. If we accept the proposition that Web documents can be incorporated into library collections, we must also address the differences between traditional and Web documents. Traditional material, once published is distributed and multiple owners hold the ownership of at least the physical entity and sometimes the intellectual content. Web documents reside on single servers and sometimes on a limited number of mirrors. Their continued existence can be decided by Web authors or Webmasters. Once a Web site is removed from its server, unless it is cached or archived it is gone.

It is true that the media that traditional materials are produced on deteriorate over time. It is also possible for societies to seek to eliminate whole bodies of literature either from neglect or active eradication. The loss of the collection from the Library at Alexandria is an example of both, but perhaps as Ray Bradbury's novel *Fahrenheit 451* illustrates, we are reluctant to tolerate the loss of an intellectual heritage. Conceivably, Web documents are not subject to physical deterioration but they are subject to the technological.

There is an even more important difference between Web documents and traditional ones. As we have seen, once an author or publisher ceases to support a traditional publication, so long as copies are maintained in libraries, the document continues to exist and more often than not continues to be available for use. This is not so with the Web. Moreover, when a Web document is edited or updated, the edited version replaces and erases the earlier "edition." New traditional editions may update older ones but they do not erase them.

Creating Web document archives and caches can solve some of these issues. A Web cache may be considered a very short term Web archive. In caching, Web documents are stored then accessed at some location "closer to" or faster for the end user than the original host is. Caching creates ethical, legal, as well as timeliness concerns (Tewkesbury, 1998). Archiving raises additional problems including what to archive and how often should the archive be updated. Storage and retrieval can quickly become a major problem. For more on Web archives see Kahle (1997) and on digital collections in general the *Report of the Task Force on Archiving of Digital Information* (Commission on Preservation and Access and The Research Libraries Group, Inc. 1996).

We are concerned here with the library of native documents rather than with the library of cached documents. The literature suggests for example that change and attrition effect information retrieval. Changing pages are the more frequently visited pages. Based on a limited sample, *Douglis, et al.* (1997) have addressed caching efficacies and have explored re-access patterns to previously accessed Web material. Among their conclusions are: "…the most frequently accessed resources have the shortest intervals between modification." In a similar vein, Tauscher and Greenberg (1997) find that "… 58% of an individual's [page browsing habits] are revisits, and that users continually

add new web pages into their repertoire of visited pages. People tend to revisit pages just visited, access only a few pages frequently, browse in very small clusters of related pages, and generate only short sequences of repeated URL paths."

The organization of Web sites may help predict document longevity. Koehler (1999a) has shown that Web pages located lower (level two plus) on the server file structures are more than twice as likely to go comatose but are less than twice as likely to change than are documents at higher levels (levels zero and one). Web pages located at the server-level domain (http://aaa.bbb.ccc) are zero level pages, those one file removed (http://aaa.bbb.ccc/nnn) are at the first level, and so on. This is not particularly surprising in that many zero and first level pages are navigational and will necessarily change as the Web site as a whole changes, while those pages located deeper on the file structure are more likely to be "content" in nature.

Web sites and Web pages undergo content change and attrition. While content and structural changes can and have been measured (Koehler, 1999a), this paper limits its focus to Web site and Web page persistence, or what the University of Waterloo's Scholarly Societies Project has defined as URL stability (1999). Persistence is the most fundamental form of change. Either a document exists or it does not. Moreover, in paraphrase of Hamlet's exquisite query, not only is it a question of being and of not being, it is also a question of being again. No library, digital or traditional, can adequately manage its collection if the very existence of that collection is either in flux or in question.

Research design

Data Collection

It has been argued that Web site persistence defines Web site stability (University of Waterloo 1999). The data presented in this paper were first collected in December 1996. Web page data are collected on a weekly basis using FlashSite 1.01. Web page data were also collected beginning in December 1996 and are reported here through January 21, 1999, or 106 weeks. FlashSite is automated Web page maintenance software that captures data on Web page response and if present, the size of the Web page in kilobytes (byte-weight), and additions or deletions of hypertext links to other Web objects. The sample is a randomly selected group of 361 then viable Web pages generated from the WebCrawler random URL generator. The WebCrawler program is no longer available. The sample was stratified according to the distribution of top-level domains as reported by NetCraft for December 1996.

Web Site Sample

The Web site sample was derived from the Web page sample. Saving Web page URLs to the site home or index page created the 344 Web site sample. The Web site was then mapped using WebAnalyzer 2.0, a Web site analytic and diagnostic program. WebAnalyzer data include number and types of Web objects, the number of hypertext link levels on the site, and the total byte-weight of text and graphic objects. The original sample consisted of 344 Web sites; reduced from 361 because of attrition or because of various barriers to full accesses to the site. Web site data have been collected three additional times: in July 1997, February 1998, and October 1998. Both programs used in this study, WebAnalyzer and FlashSite are products of InContext.

Definition of Terms

This study is "about" Web page and site persistence and their implications for libraries of Web material. Persistence can be assessed not only by the continuing existence of Web pages and Web sites. Both Web pages and sites manifest one of three forms of persistence behavior. They may persist of a given period of time – which is to say each time the URL is queried, the specific page addressed is accessed, it "responds." The Web page may fail to respond. For purposes of this research, any Web page that fails to respond, but subsequently responds after one of two additional queries in a forty-eight hour period is considered "present." Web page data are collected on a weekly basis. A Web page is defined here as "comatose" if on the sixth and subsequent weekly queries including the most recent when data were taken, that Web page failed to resolve. The term "comatose" is chosen purposefully, for as the data which follow demonstrate, once a Web page goes comatose, this in itself is no guarantee that a Web page will not return at some time to the same URL. If a Web page were to return after the sixth collection period or five weeks, it would no longer be counted in subsequent assessments as comatose. Web pages that fail but return are

labeled "intermittent." The longest consecutive period for the sample a Web page was comatose but returned is 84 weeks. About 8.3 percent of the sample present at the January 21, 1999 collection was comatose over more than half of the two year period, although not always consecutively. Web pages that " go and come again" are said to be intermittent. After 106 weeks, each of the 361 Web pages in the sample were present, on average, 74.2 percent of the time.

Results

The data that follow describe the persistence lifecycles of Web sites and Web pages. Additional and more extensive work is required to more fully understand the dynamics of Web content change and demise, particularly as Internet and subject domains, publisher, content type, Web site type, longevity, change frequency, object mix, subject matter, and other Web site characteristics. The list might also include frequently accessed material by subject area (e.g., Spink, *et al.*, 1998); the extent of Web linkage patterns to include the number, type, and level of hypertext interconnection; language patterns (e.g., Koehler, 1996); and taxonomies (e.g., DeRose, 1989; Haas & Grams, 1998).

Web Sites

Table 1 provides data for a randomly collected sample of 344 Web sites first identified in December 1996.

Non-Intermittent Web Sites Comatose After Indicated Date (In Percent, n=344)	
Dec-96	7.8
Jul-97	2.6
Feb-98	11.3
Total in Oct-98	21.7
Intermittent Web Sites: Date(s) Not Present	
Jul-97	3.2
Jul 97/Feb 98	1.5
Feb-98	6.7
Jul 97/Oct 98	0.9
Total	12.3
Always Present All Collection Dates	
	66.0

Table 1. WebSite Attrition-Intermittence Distributions. December 1996 to October 1998

	Total Sample (In Percent, n=344)			Goners Only	
Ever Present	Comatose & Intermittent				
		Comatose	Intermittent	Comatose	Intermittent
Total	66.0	21.8	12.3	64.1	35.9
Commercial	59.3	28.6	12.1	70.3	29.7

Educational	61.2	20.9	17.9	53.8	46.2
ISO	73.6	16.0	10.4	60.6	39.4
Gov-Mil	76.2	19.0	4.8	80.0	20.0
Net-Org	60.0	27.5	12.5	68.8	31.3

Table 2. Web Site Persistence, Attrition, and Intermittence. Total and by TLD, December 1996 to October 1998

Based on the data presented in Tables 1 and 2, nearly twenty-two percent of the sample was "unstable" over a twenty-two month period, while another 12.3 percent were "partially unstable." It is clear that some Web site classes are more unstable than others are. For example, as shown in Table 2, commercial, educational, and network and organizational sites have a higher "instability" rate than the ISO or government and military sites. How these sites are "unstable" also varies. Educational sites are more intermittent than the overall norm but are less "comatose." Commercial sites, on the other hand, are much more likely to go comatose than to be intermittent. The University of Waterloo Scholarly Societies Project (1999) reports that Web site stability is related not only to the "canonical" form of the URL name, but also to the location of the Web site on the server-level domain (SLD). The closer the Web site index page address to the SLD, the more stable it is. Their study is largely limited to Web sites on the org TLD. Similar results for Web sites on all TLDs have also been reported (Koehler, 1999a). Thus persistence or stability is the first order question.

Over the almost two year period, almost two-thirds of the Web site sample responded in each of the four periods when queried. Between December 1996 and July 1997, 7.8 percent of the sample failed to respond and continued to not respond in subsequent periods. By October 1998, just more than one-fifth of the sample was considered comatose in that once a Web site failed it continued to fail. By October 1998, an additional eighth of the sample was defined as "intermittent" because they would fail but reappear at a later time. These data are presented in the second row of Table 1. For example, 3.2 percent of the total sample failed to respond in July 1997, but returned for subsequent queries. At the same time, 0.9 percent failed in July 1997 and again in October 1998.

As I have shown elsewhere (Koehler, 1999a,bk), Web sites can be categorized according to a variety of general and Web-specific criteria. Table 2 illustrates Web site attrition and intermittence behavior for the sample by top-level domains. Gov-Mil and Net-Org are combined on the Table because first they individually comprise a small proportion of the sample and second because to date they have manifested similar behavior. The term "goner" is used to describe Web sites that have failed to respond at least once. The column labeled "comatose" refers to those Web sites "gone" in October 1998. The "intermittent" column refers to those "back" in October 1998 but that had been gone during earlier collections.

There are, for the present, two classes of top-level domains (TLD): functional and geographic or gTLDs (generic TLDs) and ccTLDs (country code TLDs). Functional TLDs describe the corporate or institutional type of Web site publisher. Individual isodomains include com for commercial, edu for educational, org for organization, net for network, gov for government, and mil for military. A rarer form is int for international organization. The geographic indicate the country location of the publisher. Cc-isodomain examples include au for Australia, ca for Canada, fr for France, pe for Peru, and za for South Africa. These ccTLDs are based upon the two-character International Standards Organization standard 3166 (ISO 3166).

Table 2 documents attrition-intermittence behavior for TLD groups over a twenty-two month period beginning in December 1996. The column labeled "Total Sample" provides data for Web sites that persisted over the entire sample period as well intermittent and comatose sites. The second column, "Goners Only" provides the ratio by TLD for the intermittent-comatose portion of each sub-sample.

Web Pages

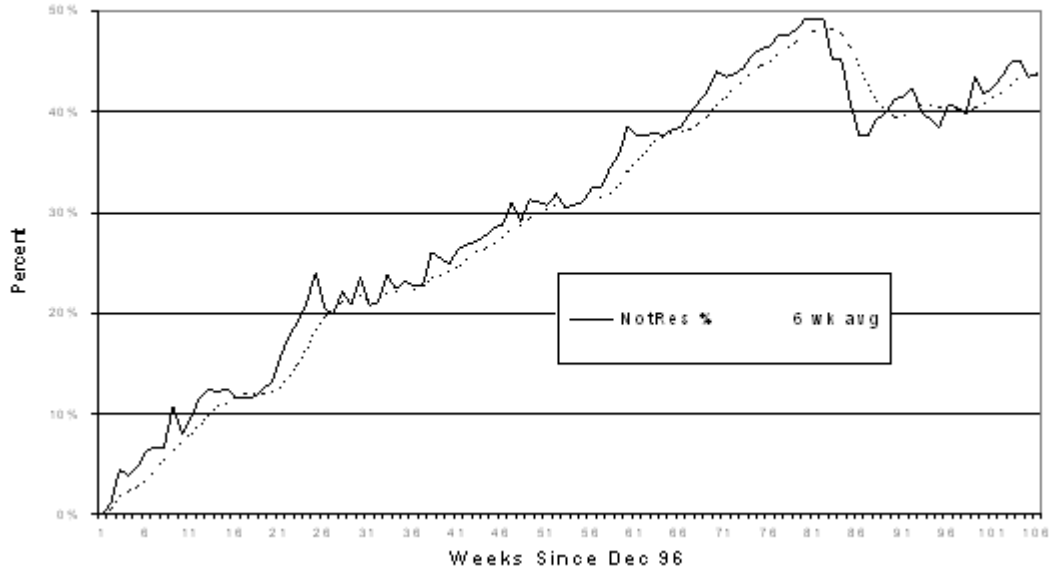
Web pages also exhibit attrition and intermittence behavior. However, when these data were calculated by TLD, Web pages manifested variations: Network (68.4 percent) and ccTLD (71.2 percent) Web pages were present, on average, less often than the sample mean; while commercial (74.8 percent) and educational (75.4 percent) approximated the sample mean; and government (78.1 percent), military (78.4 percent) and organizational (81.3 percent) Web pages were present more frequently than the average. During the same period, 42 percent of the sample met the comatoseness test. Of the 58 percent "present," 68 percent had never been comatose – or "gone" for

six or more consecutive collection periods. However, of those which had never been comatose, just less than half (47 percent) had been intermittent (or "gone" less than six consecutive periods) at least once over the two year period. In the end, only 21 percent of the Web page sample resolved each and every week.

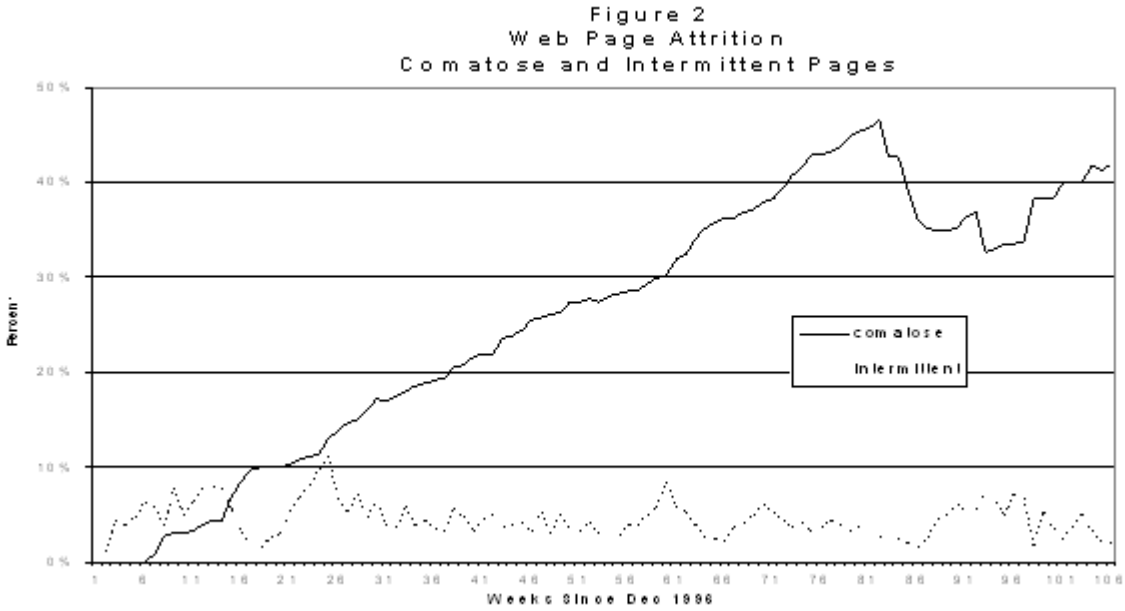
Web pages present on January 21, 1999, were comatose or missing at least six consecutive periods, an average 9.5 periods. If those Web pages present that have never been comatose are excluded, the average present Web page that has ever been comatose has been gone 29.6 periods. Figures 1 and 2 illustrate the great variability in Web page comatoseness over time.

The solid trace in Figure 1 plots the rate at which Web pages cease to respond to weekly queries beginning in December 1996. The dashed trace plots the same data averaged over the "six-data point comatoseness" period done to smooth sample based weekly data anomalies. This represents two related phenomena. Web authors frequently update and edit their Web sites. They may prune Web pages from the site and this results in one form of comatoseness. Web sites, as a whole, may also cease to respond for any number of reasons. The term comatose page is preferred because, as is shown in Figure 2, any given time about five percent of the sample is intermittent and will return.

Figure 1
Web Page Attrition



Web pages (and Web sites) are sometimes resurrected, and the same Internet address or URL is used to present sometimes similar and sometimes very different information. As is shown in Figure 1, significant resurrection activity took place from week 82 to week 87 (early August to mid September 1998). By August 1998 nearly half the sample failed to resolve during the week 82 collection and as is shown in Figure 2 nearly as many were defined as comatose. However, by mid-September, nearly ten percent of the "missing" sample had returned.



Attrition and digital libraries

Two forces effect library collection development from the Web. In order to provide patrons with both broad and exhaustive resources libraries will necessarily have to turn to digital and Web resources to expand and augment their collections. In the end, digital and hybrid libraries, I believe, will have to incorporate native Web documents into their collections to a degree far greater than is now the practice.

The Web and for that matter traditional publishing are in constant flux. New Web documents are constantly being added to the available resource pool, just as new print and digital materials are. This offers no original challenge to libraries. The original challenge to libraries brought by the Web lies in its ephemeral nature. Both Web pages and Web sites undergo metamorphosis, attrition, and return. Any library that seeks to incorporate Web materials into its collection must take into account this dynamic or the quality of the bibliographic representation as well as of the collection itself will rapidly erode. If indeed the half lives of Web sites and Web pages are about two years, information managers must pay attention to the process. For by the end of two years, half of the original collection will have gone, and by the end of ten years, only 3.125 percent of the collection may remain.

The good news is that these numbers probably represent a worst case scenario. Collection developers can follow the model offered by the Scholarly Societies Project at the University of Waterloo or as this paper suggests and collect only those Web sites meeting certain criteria, including probable longevity. Nevertheless and no matter how rigid the selection policy, Web-derived collections are far less persistent than any other kind. Not only must collection developers identify new Web material, they must also cull dead links and catalogue entries and replace them with equivalent resources and new catalogueing.

A second alternative is to further our understanding of the lifecycle of Web documents and to use that understanding to better control for the impacts and vicissitudes inherent in that cycle. It is possible to predict which pages and sites are likely to disappear or to be intermittent. This information itself can be incorporated as part of the bibliographic record (Koehler, 1999b).

Document death is never trivial, for in the absence of an archive, the information contained in that document is lost. Comatoseness is not necessarily as tragic as document death in that some documents are intermittent and do return. This may not be unique to the Web, for as Weinberg (1999) argues, librarians have had extensive experience managing changing technology, changing media, and controlling for document demise. The document lifecycle is far more prevalent on the Web than for print; it is in fact an accepted norm. If this lifecycle is both ubiquitous and normal, those who seek to manage Web content must develop new systems and adapt old standards that specifically address it.

To manage the lifecycle, the digital library community has begun to address the dynamic and the impact that change has on the quality of collections; indexing, search, and retrieval algorithms, end user and intermediary behavior;

document storage, caching, and archiving; and other related issues. I have suggested that lifecycle data can be captured for individual Web sites and pages and used to advantage for automated indexing and catalogueing (Koehler, 1999b). We have, however, just begun to identify the advantages and disadvantages inherent in Web document metamorphoses.

References

- Abramson, A. (1998). Monitoring and evaluating use of the World Wide Web in an academic library. Information Access in the Global Information Economy. *Proceedings of the 61st Annual Meeting of the American Society for Information Science 35*, Pittsburgh, PA, October 25-29. Medford, NJ: Information Today, Inc.: 315-326.
- Almind, T. & Ingwersen, P. (1997). Informatic analyses on the World Wide Web: methodological approaches to "Webometrics." *Journal of Documentation* **53**, 404-26.
- Arnold, K. (1994). The Electronic Librarian Is a Verb/The Electronic Library Is Not a Sentence. A Lecture Delivered at the New York Public Library. The Gilbert A. Cam Memorial Lecture Series October 14, 1994. *Biblion*.
- Auer, N. (1998). Bibliography on evaluating Internet resources. Last updated: October 16, 1998.
- Bailey, C.W. (1999). *Scholarly Electronic Publishing Bibliography*. Version 24: 4/1/99.
- Banerjee, K. (1997). Describing remote electronic documents in the online catalogue: Current issues. *Cataloguing & Classification Quarterly* **25** (1).
- Baron, L., Tague-Sutcliffe, J. & Kinnucan, M. (1996). Labeled, typed links as cue when reading hypertext documents. *Journal of the American Society for Information Science* **47**, 896-908.
- Borenstein, N. (1991). *Programming as if people mattered: friendly programs, software engineering, and other noble delusions*. Princeton: Princeton University Press.
- Brake, D. (1997). Lost in cyberspace. *New Scientist* **154** (2008): 12-3.
- Bray, T. (1996). Measuring the Web. *Fifth International World Wide Web Conference*. May 6-10, 1996, Paris, France.
- Caplan, P. (1994). Controlling E-Journals: the Internet Resources Project, catalogueing guidelines, and USMARC. *The Serials Librarian* **24**, 103–111.
- Chen, C-c. (1998). Global digital library: Can the technology havenots claim a place in cyberspace? In Ching-chih Chen, ed., *Proceedings NIT '98: 10th International Conference New Information Technology*, Hanoi, Vietnam, March 24-26, 1998. West Newton, MA: MicroUse Information, 1998: 9-18.
- Chu, H. (1997). Hyperlinks: How well do they represent the intellectual content of digital collections? *Proceedings of the 60th Annual Meeting of the American Society for Information Science, vol. 34. Digital Collections: Implications for Users, Funders, Developers and Maintainers.* Washington, DC, November 1-6, 1997. Medford, NJ: Information Today, 361-69.
- Ciolek, T.M. (1996). The six quests for the electronic grail: Current approaches to information quality in WWW resources. *Revue Informatique et Statistique dans les Sciences Humaines* (1-4), 45-71.
- Commission on Preservation and Access and The Research Libraries Group, Inc. (1996). *Preserving Digital Information*. Report of the Task Force on Archiving of Digital Information.
- DeRose, S. (1989). Expanding the notion of links. *Proceedings of Hypertext '89*. ACM: 249-57.
- Desai, B. (1997). Supporting discovery in virtual libraries. *Journal of the American Society for Information Science* **48**, 190-204.
- Digital Library Information and Resources (1999).
- Dillon, M., Jul, E., Burge, M. & Hickey, C. (1993). *Assessing information on the Internet: Toward providing library services for computer-mediated communication*. Dublin, OH: OCLC Office of Research.
- Douglis, F. (n.d.). Internet difference engine research.
- Douglis, F., Ball, T., Chen, Y-F. & Koutsofios, E. (1996). WebGUIDE: Querying and Navigating Changes in Web Repositories. *Fifth International World Wide Web Conference*. May 6-10, 1996 - Paris, France.
- Douglis, F., Feldmann, A., Krishnamurthy, B. & Mogul, J. (1997). Rate of change and other metrics: A live study of the World Wide Web. *USENIX Symposium on Internet Technologies and Systems*, December 1997: 147-158.
- Dublin Core Metadata (Last updated November 2, 1997).
- Dublin Core Metadata Initiative (Last updated May 16, 1999).
- Ellis, D., Ford, N. & Furner, J. (1998). In search of the unknown user: Indexing, hypertext and the World Wide Web. *Journal of Documentation* **54**, 28-47.

- Forrester, M. (1995). Indexing in hypertext environments: the role of user models. *The Indexer* **19**, 249-256.
- Griffiths, J-M. (1998). Why the Web is not a library. B.L. Hawkins and P. Battin, eds. *The Mirage of Continuity: Reconfiguring Academic Information Resources for the Twenty-First Century*. Washington, DC: Council on Library and Information Resources.
- Haas, S. & Grams, E. (1998). A link taxonomy for Web pages. Information Access in the Global Information Economy. *Proceedings of the 61st Annual Meeting of the American Society for Information Science 35*, Pittsburgh, PA, October 25-29. Medford, NJ: Information Today, Inc.: 485-95.
- Hsieh-Yee, I. ([1996]). Modifying catalogueing practice and OCLC infrastructure for effective organization of Internet resources. OCLC Internet catalogueing Project Colloquium Position Paper.
- Johnston, C. (1998). Electronic Technology and Its Impact on Libraries. *Journal of Librarianship and Information Science* 30 (1): 7-24.`
- Jul, E., Childress, E. & Miller, E. (1997). 42: Don't Panic, It's a Common Disaster and 42: Now That We Know the Answer, What are the Questions? *Journal of Internet catalogueing* **1** (3).
- Kahle, B. (1997). Preserving the Internet. *Scientific American* **276** (3), 82-83.
- Kaiser, J., ed., (1998). New search strategy untangles the Web. *Science* **280** (5364), 647
- Khan, K. & Craig Locatis, C. (1998). Searching through cyberspace: The effects of link display and link density on information retrieval from hypertext on the World Wide Web. *Journal of the American Society for Information Science* **49**, 176-82.
- Koehler, W. (1996). A descriptive analysis of Web documents and demographics. *Proceedings NIT '96: 9th International Conference New Information Technology*, Pretoria, South Africa, November 11-14, 1996. West Newton, MA: MicroUse Information: 159-170.
- Koehler, W. (1998). The librarianship of the Web: Options and opportunities managing transitory materials. In Ching-chi Chen, ed., *Proceedings NIT '98: 10th International Conference New Information Technology*, Hanoi, Vietnam, March 24-26, 1998. West Newton, MA: MicroUse Information, 1998: 97-106.
- Koehler, W. (1999a). An analysis of Web page and Web site constancy and permanence. Forthcoming *Journal of the American Society for Information Science.*
- Koehler, W. (1999b). Classifying Websites and Webpages: The use of metrics and URL characteristics as markers. Forthcoming *Journal of Librarianship and Information Science.*
- Koehler, W. & Mincey, D. (1996). FirstSearch and NetFirst Web and Dialup Access: Plus Ça Change, Plus C'est La Même Chose, *Searcher* **4**, (6), 24-8.
- Lawrence, S. & Giles, C.L. (1998). Searching the World Wide Web. *Science* **280** (5360), 98-100.
- Lesk, M. (1996). Libraries and the Web: 1995. *Libraries and Information World Wide*.
- Lesk, M. (1997). *Practical digital libraries: Books, bytes, and bucks*. San Francisco: Morgan, Kaufman.
- McDonnell, J., Koehler, W. & Carroll, B. (1997). Automating the dynamic development and maintenance of a distributed digital collection: The Area Studies Digital Library (ASDL), *Proceedings of the 60th Annual Meeting of the American Society for Information Science, vol. 34. Digital Collections: Implications for Users, Funders, Developers and Maintainers.* Washington, DC, November 1-6, 1997. Medford, NJ: Information Today: 244-61.
- McDonnell, J., Koehler, W. & Carroll, B. (1999). Cataloging challenges in an area studies virtual library catalog (ASVLC): Results of a case study. Forthcoming in *Journal of Internet Cataloging*, **2**(2), 15-42.
- Nordic Metadata Projects (updated 25 June 1998).
- Oder, N. (1998). catalogueing the Net: Can we do it? *Library Journal* (October 1, 1998): 47-51.
- Olson, N., ed. (1997*). Cataloging Internet resources. A manual and practical guide*. 2ed.
- Pattie, L-Y. & Cox, B.J., eds. (1996). *Electronic resources: selection and bibliographic control*. NY: Hayworth Press.
- Pinfield, S., Eaton, J., Edwards, C., Russell, R. & Wissenburg, A. (1998). Realizing the hybrid library. *D-Lib Magazine* (October).
- Prescott, Andrew (n.d.) The Digital Library in theory and practice: A historian's view.
- Janet Reno, Attorney General of the United States, et al., Appellants v. American Civil Liberties Union et al. No. 96-511, Supreme Court of the United States, 117 S. Ct. 2329; 1997 U.S. Lexis 4037.
- Shvartsman, M. (1998). Creation of systematic cataloguing of Russian resources on the Internet. *Libraries and Associations in the Transient World: New Technologies and New Forms of Cooperation, Proceedings "Crimea98" International Conference,* Sudak, Ukraine, June 6-14, 1998. [Moscow]: NPLS&T, v. 1, 238-9.
- Spink, A., Bateman, J. & Jansen, B. (1998). Searching heterogeneous collections on the Web: behavior of Excite users. *Information Research* **4** (2).
- Tauscher, L. & Greenberg, S. (1997) Revisitation patterns in World Wide Web navigation. *Proceedings CHI*

97 Conference on Human Factors in Computing Systems Atlanta, Georgia, March 22-27, 1997.
- Tennant, R. (1998). The art and science of digital bibliography. *Library Journal digital* (October 15, 1998).
- Tewkesbury, R. (1998). Is the Internet heading for a cache crunch? *On the Internet* **4**(1), 17-22.
- Tilman, Hope. (1998). Evaluating quality on the Net.
- Tomaivolo, N. & Packer, J. (1996). An analysis of Internet search engines: Assessment of over 200 search queries. *Computers in Libraries* **16** (6): 58-62.
- Travica, B. & Olson, R. Web as global virtual library: Usability of business sites in East and Central Europe. . Information Access in the Global Information Economy. *Proceedings of the 61st Annual Meeting of the American Society for Information Science 35*, Pittsburgh, PA, October 25-29. Medford, NJ: Information Today, Inc.: 227-42.
- UKOLN Metadata Dublin Core registries (Last updated November 27, 1997).
- University of Waterloo, Scholarly Societies Project (Last updated January 27, 1999).
- Weinberg, B. (1999). Improved Internet access: Guidance from research on indexing and classification. *Bulletin of the American Society for Information Science* **25** (2): 26-9.
- Wells, H.G. (1938). *World brain*. Garden City, NY: Doubleday, Doran, and Co.
- Woodruff, A., Aoki, P., Brewer, E., Gauthier, P. & Rowe, L. (1996). An Investigation of Documents from the World Wide Web. Fifth International World Wide Web Conference May 6-10, 1996, Paris, France.

Note

Since its publication in *Information Research* this paper has been translated into Chinese and published in *New Technology of Library and Information Service*, **no. 6**, 2000, 75-79

How to cite this paper:

Koehler, Wallace (1999) "Digital libraries and World Wide Web sites and page persistence." *Information Research*, **4**(4) Available at: http://informationr.net/ir/4-4/paper60.html

Contents **Home**