An Analysis of Web Page and Web Site Constancy and Permanence

Wallace Koehler
School of Library and Information Studies, University of Oklahoma, Norman, OK 73019-0528.
E-mail: wkoehler@ou.edu

We recognize that documents on the World Wide Web are ephemeral and changing. We also recognize that Web documents can be categorized along a number of dimensions, including "publisher," size, object mix, as well as purpose, meaning, and content. This study is first a preliminary exploration into Web page and Web site mortality rates. It then considers two types of change: Content and structural. Finally, the study is concerned with understanding those constancy and permanence phenomena for different Web document classes. It is suggested that, from the perspective of information maintenance and retrieval, the WWW does not represent revolutionary change. In fact, in some ways the Web is a less sophisticated form than traditional publication practices. Finally, this study explores the "short memory" and "mind changing" of the World Wide Web.

1.0. Introduction

We are, it is alleged, in the middle of an information revolution. We can communicate, access databases, and publish almost anything we wish at relatively low costs and at ever increasing speeds. The Internet and its "publishing arm," the World Wide Web, are important components in that communication process.

If we are in an information revolution, it is an odd revolution. Without question, we can send, access, manipulate, and publish information or data over longer distances with almost no barriers and virtually at the speed of light. This ease of publication and distribution does not constitute a "Kuhnian" paradigm shift (Kuhn, 1970) in our understanding of information, informatics, information manipulation, or information transfer. It does not represent a Marxian redistribution of information resources or control of those resources from one group to another. In fact, some worry that new divides between information haves and have-nots will result (e.g., Sandar & Ravitz, 1996). It will not result in a Weberian redefinition of institutions for information creation and distribution, although some change has already occurred and more is inevitable (e.g., Lesk, 1997; Shuman, 1997).

The Web does not represent a new order of magnitude in information quality, although it may in quantity. It is, instead, the next step in a long-established evolutionary process with social, political, economic, and institutional overtones in information authoring, analysis, distribution, and retrieval (e.g., Lubar, 1993; Lynch, 1997; Stefik, 1996). The Internet and its information content represent an extension of "Price's Law," expanding Price's insights from the exponential growth of the scientific literature (Price, 1963, 1965, 1976) to a wider literature.

If we are in an information revolution, it is not a revolution of content or of access. It is, instead a revolution of process (Snyder, 1997). We recognize that Web documents are not the same thing as published and immutable works. Nor do they disappear the very moment they are uttered or broadcast. The WWW represents a third model that coexists between the recorded and the unrecorded.

This new model of information transfer and communication raises both philosophical and practical questions. Human communication can be thought of in two formats: Ephemeral and "permanent" (e.g., Dunbar, 1996). Web documents may be considered a form of human communication that lie somewhere between ephemera and permanent (Feldman, 1997; Koehler, 1998). The World Wide Web is itself a publishing anomaly. The uttered word is meant to be consumed immediately. It is ephemeral; once spoken it is lost. The recorded word is designed to be preserved; it persists. These intended lifetimes drive useful lifetimes, formats, update schedules, retrieval systems, and other distinctions.

Web pages and sites may be recorded, but most either disappear or their content is modified and overwritten, leaving no trace of the earlier document. If recorded (cached), the cached file may differ from the changed native file (Tewksbury, 1998). Web document storage and maintenance are centralized in a distributed system. Web authors and Webmasters own and control the single, or a very small

Received February 4, 1998; accepted March 30, 1998.

number of, definitive Web document copies. While end users may access and cache those documents, the authority inherent in distributed print material does not adhere to cached Web documents (Tewksbury, 1998). Web documents are modified at will and are often changed frequently. Limited copies exist and are centrally owned. In that, Web documents are more like ephemera than published works. Yet, once uttered they do have a "lifetime," they can be copied and recorded, and those copies can be stored and accessed (Feldman, 1997; Kahle, 1997).

The WWW may represent an intermediate form between recorded and unrecorded communications and information transfer. Because it is a new medium, we have not yet fully identified the dynamics of its behavior. This article begins to address the permanence and constancy cycles of the Web as part of an effort to understand that dynamic. It is suggested that permanence and constancy can be understood in the context of Web domains, and Web page and site structures. By charting Web page and Web site constancy and permanence behavior, this article begins to map an understanding of the WWW as a third model.

2.0. Statement of the Problem

This article addresses two aspects of Web site and Web page behavior: The permanence of Web pages and sites, and the constancy of those Web pages and sites. We all recognize the transitory nature of Web documents. How transitory are they? How often do they change, what changes, and does it matter? This article focuses on three important questions:

1. How permanent are Web pages, Web sites, or server-level domains? What is the death rate for each? What is the resurrection rate for each? How often do they move?
2. How constant are Web pages and sites?
3. Do different types of Web pages, Web sites, and domains behave differently? Web pages and Web sites may be distinguished by several criteria. Two are employed here: First, those that can be determined from an examination of the URL; and second, those that can be derived from quantitative measures of Web sites and pages.

These questions have not yet been often addressed in the literature. I have offered some preliminary findings based on this and related research (Koehler, 1996, 1997b). Chankhunthod, Danzig, Neerdeals, Schwartz, and Worrell (1995; http://excalabur.usc.edu/cache-html/subsectionstar3_4_0_1.html) report that the mean lifetime of all WWW objects was over a 3-month period in 1995, 44 days, and that html text and image objects are limited to lifetimes of 75 and 107 days, respectively. Their work reported lifetimes for specific Web objects rather than for collections of Web objects—which are what Web pages and Web sites are. For the information scientist, librarian, and others concerned with the transmission or transfer of information from an author to an end user, the lifetimes of Web pages and Web sites are more useful in managing that transfer than the lifetimes of individual Web objects.

The Scholarly Societies Project at the University of Waterloo maintains a virtual library of professional society Web sites. The project editor, Jim Parrot, has developed a URL-Stability Index. The Index is applied to each discipline and the value indicates the URL stability of the Web pages within each professional group. The Index is built on the assumption that "canonical" URLs, those which contain the society name and are located at the server-level domain are unlikely to disappear (www.orgname.org). There are two less stable forms. Those URLs not opening at the SLD are less stable than those that do: www.orgname.org/index is less stable than www.orgname.org. Finally, those URLs that are not canonical, that is, do not contain the organization name within the domain name, are the least stable: www.univ.edu/~orgname. The University of Waterloo Index seeks to predict URL death, but it does not address content change (University of Waterloo, 1997).

As the Scholarly Societies Project demonstrates and is argued here, different types of Web entities change and disappear differently. Web site and page taxonomies were defined and developed in the thesis this article grew from (Koehler, 1997c). Web entities can be classified using a variety of markers. One such set of markers consists of the elements or fragments incorporated into URLs. These markers include domain names (for a discussion of these practices, see Koehler & Barnett, 1998), directory structure of the site, semiotic tag patterns (Urgo, 1996), unusual ports, and the use of tildes to attach documents to Web sites. A second set of markers can be derived from analyses of Web site and page size, and object counts and mixes (Koehler, 1997c). This second set also includes hypertext link patterns (e.g., Chu, 1997; Khan & Locatis, 1998) and the relationship of any given page to other members of the site set.

These markers can be used to build Web page and site classification schemes. These, in turn, can be utilized to help understand and identify Web site and page constancy and permanence behaviors. A taxonomy of Web sites, based on object distribution and number, has been developed (Koehler, 1997c), and the constancy and permanence of these Web site types are reported here.

There is a wide range of possible third model descriptive and predictive analyses that can be performed. I suggest that the first order questions to be addressed are:

1. Web sites are inherently more permanent than specific Web pages, since Web sites consist of Web pages. Web pages can depart without effecting Web site demise. Different Web page types will manifest different permanence and constancy characteristics. Because of the changing nature of what they point to, navigational pages change on aggregate more often than do "content" or substance pages. However, content pages are less permanent than navigation material. Content pages are less permanent than the navigation because navigation pages must be retained to point to new content, while content pages need not point to navigation.

2. Web documents behave differently according to publisher type. Publishers can be identified by TLD, sometimes 2LD. Where publisher identification is not explicitly provided, it can often be inferred from the URL. The following distinctions among publisher types can be tested:
   a. Commercial pages change more often than others do. Commercial sites are smaller. Their purpose is to sell and promote products. Web pages will turn over as products/information/prices change. The growth rate of individual com sites will be minimal.
   b. Net and edu sites share certain similarities, e.g., the use of tildes[1] (~) to mark levels of discontinuity. Edus offer Web sites of record for academic and educational publishing, as well as for student experimentation. The new will tend to displace the old less often. Therefore, edu sites will increase in size over time. Net sites will grow through accretion. As new subscribers add content, size will increase. Net pages will turn over more rapidly than edus since nets are not "repositories of record."
   c. The ISO 3166 domains are a mixture of types. Functional domains publishers inferred from ISO-tagged domains behave like their functional counterparts.
   d. Orgs are more stable than coms, but exhibit similar attributes. Orgs "sell" not-for-profits.
   e. Gov and mil explicit and inferred domains are the most permanent and constant of the domains. An explicit domain carries the appropriate functional TLD or 2LD, and takes the form xxx.mil or gov.xx. An inferred domain is one where the functional domain can be inferred from an examination of the URL. Examples include the following Web sites on the educational inferred domain: ubonn.de, Bonn University; mcgill.ca, McGill University; and u-lyons.fr, University of Lyons. Content is often "record." Record documents are most permanent and constant on the WWW.
3. Web page and site structures can be captured using software assisted quantitative methods. They describe several Web attributes that contribute to our understanding of Web site, and page permanence and constancy. Web site and page structures include the following:
   a. Size. Web sites consist of one or more Web pages. Web pages are files that contain some number of Web objects. The number of objects and hypertext links on a page can be counted. In turn, the number of pages or files can be also be counted. Size can also be determined as "byte-weight," that is, the number of bytes of which each object, page, or site consists.
   b. Web site structures. Web sites are organized in two ways: Hypertext links and directory hierarchy. Web sites can be mapped using either approach with very different results. The locus of any given Web page can also be mapped with either technique. The hypertext relationships of set members indicate not only the Web authors' subjective ordering schemes, they also provide an indication of site "depth." Site depth is one element in developing site size, complexity, and density measures.
   c. Web page loci. Web pages are located within Web sites by both hypertext placement and on the directory hierarchy. The location of any given Web page on the directory hierarchy may provide clues to its site function. Navigational pages are located closer to the top of any given hierarchy, while content pages are usually found further down.

3.0. Web Growth and Change

The increase in the number of Web sites, Web pages, Web authors, Web publishers, and in the transmission of that information to larger and more disperse audiences parallels the increases in traditional recorded media. The Internet also parallels the use of sophisticated electronic technology to distribute both recorded and ephemeral communications.

3.1. General Indicators of Internet Growth

Internet changes and with them, World Wide Web changes have been the subject of many studies and statistical reports. Internet histories have been published that recount not only the technologies and events leading up to the creation of the Internet, but also its growth from a defense oriented academic network to the commercial enterprise that it has become (Rough Guide to the Internet, 1997; Zakon, 1997). Lawrence Landweber (1997) has until recently published an international interconnectivity map (the "purple map") on the back inside cover of each number of On the Internet, one of the Internet Society publications. The purple maps have shown an ever-increasing E-mail, bitnet, and Internet interconnectivity on a country-by-country basis.

Other WWW measures are increasing. NetWizards (1998) publishes a periodic survey of the number of Internet top-level domains and hosts from 1993 to the present. According to them, the number of domains increased from 21,000 in January 1993 to 4.3 million in July 1997. Similarly, the number of hosts increased from 1.3 million to 19.5 million over the same period.

Matrix Information & Directory Services (1997) publishes Internet user, systems, network, demographic, and interconnectivity data. This is a proprietary service with charges for most information. Their data demonstrate increases in telecommunications infrastructure and message traffic, again from 1993. The number of WWW bytes transmitted increased from less than $1 * 10^5$ at the end of 1992 to more than $1.5 * 10^{12}$ in early 1977. The number of WWW packets increased from less than 300,000 to over $1 * 10^7$, again over the same period.

3.2. Managing Constancy and Permanence

Constancy and permanence issues have generated some attention. While statistics for WWW constancy and perma-

[1] Tildes are often required UNIX punctuation. Thus, the use of the tilde does not necessarily mark a point of discontinuity. UNIX systems are most often employed by academic and network users, hence the predominance of tildes on edu and net domains.

nence have rarely been collected, the transitory or nonpermanent nature of URLs has been the focus of much attention. The Internet Engineering Task Force (IETF) is exploring other URxs to augment URLs. These are URNs (uniform resource names), URIs (uniform resource identifiers), and URCs (uniform resource characteristics). These have been proposed as possible solutions to the "unstable" character of URLs (World Wide Web Consortium, 1997; WSC Architecture Domain, 1997).

PURLs (persistent URLs) have been offered as alternative solution to transitory URLs and the IETF proposals by OCLC. PURLs, rather than pointing directly to a WWW resource, point instead to an intermediate inventory. The resolution service would translate the PURL to the then functional URL to provide access to the Web document (OCLC, 1997).

The URx can address WWW document changes by creating a unique address for each "edition" or metamorphosis of a Web page. These options can also offer a solution to Web page and Web site demise by archiving each iteration or change that the Web entity experiences. PURLs can also point to archived material.

Interesting as the IETF and OCLC proposals are, neither can be applied without an underlying and permanent collection of Web documents. Perhaps the most ambitious and necessary solution to the problem of URL inconstancy and impermanence are archives (Feldman, 1997), including the comprehensive archive proposed by Kahle (1997). By taking a series of WWW "snapshots" and preserving those snapshots, the WWW as it existed at any given time can be preserved and accessed in that "frozen" state. Many questions remain for archive implementation. Will everything on the WWW be archived and how often will the collection be reiterated? How will superceded documents be tagged or cataloged in the archive? What are the hardware and software requirements of the archive?

Based on data presented below and the most recent number of sites reported by NetWizards, the WWW is a "big place." If there were 1.2 million sites and each site averaged 5.3 million bytes (exclusive of audios and videos), the size of the Web weighed at a minimum of $6 * 10^6$ gigabytes in August 1997. If an archive were to collect the entire Web on each of its daily collection passes, it could grow to more than $2.1 * 10^9$ gigabytes at the end of 1 year. Storage alone would be staggering. This research may offer an alternative model to any fixed collection schedule by establishing the probable frequencies of change for various classes of Web documents.

4.0. Definition of Terms

One purpose of this article is to report the findings of a 1-year Web document longitudinal study. Web documents include both Web sites and Web pages. Web pages are defined here as collections of Internet objects (text, graphics, E-mail, videos, various scripts, etc.) that can be navigated without recourse to hypertext linkages. They are, in effect, Web documents that can be scrolled through. A single Web page may be article length or longer, or it may consist of nothing. Web sites are collections of one or more Web pages that share some common theme or organizing principle. Web sites connect their component Web pages through hypertext links. Individual pages may be part of other Web sites, authored by others, and be hosted on any number of servers. A Web site often encompasses an entire server-level domain (SLD), but it need not. Servers can host more than one Web site at the same SLD, and these are distinguished at implicit or explicit "points of discontinuity."

Web pages and Web sites exhibit two related types of longevity behavior: Constancy and permanence. Constancy measures the rate at which Web documents are changed in any way over time, a rate labeled "omega" here. Almost without exception, over the period of a year, all Web documents are inconstant. This is only slightly more true of complex Web sites than of single Web pages. Web document constancy does not measure the importance, magnitude, complexity, or degree of change. It is binary; either change occurred or it did not.

Permanence measures the probability that Web documents will carry the same URL over time (but not necessarily to the same Internet Protocol number) or that if "moved" to a different URL, a resolvable forwarding address is provided. There is a permanence variant: Intermittence. Intermittent URLs are those Web documents that fail to respond or resolve at any given time, but return. Web documents can therefore be considered "always present," sometimes intermittent, or "comatose." A comatose Web document is one that fails to resolve the most recent query, and has failed for six or more consecutive weekly queries, including the most recent. "Comatose" is preferred to "dead" because, though infrequent, Web documents, long comatose, sometimes return.

Web pages and Web sites exhibit two related types of longevity behavior: Constancy and permanence. Constancy measures the rate at which Web documents are changed in any way over time, a rate labeled "omega" here. Almost without exception, over the period of a year, all Web documents are inconstant. This is only slightly more true of complex Web sites than of single Web pages. Web document constancy does not measure the importance, magnitude, complexity, or degree of change. It is binary; either change occurred or it did not.

Permanence measures the probability that Web documents will carry the same URL over time (but not necessarily to the same Internet Protocol number) or that if "moved" to a different URL, a resolvable forwarding address is provided. There is a permanence variant: Intermittence. Intermittent URLs are those Web documents that fail to respond or resolve at any given time, but return. Web documents can therefore be considered "always present," sometimes intermittent, or "comatose." A comatose Web document is one that fails to resolve the most recent query, and has failed for six or more consecutive weekly queries,

including the most recent. "Comatose" is preferred to "dead" because, though infrequent, Web documents, long comatose, sometimes return.

Web pages can be defined by the functions they perform (McDonnell, Koehler, & Carroll, 1997). Two functional typologies have been described: Architectural and directory. Architectural definitions describe Web page purpose within the Web site. Web pages may define the purpose, they may provide navigation, or they may offer content. Most pages are combination forms. Most homepages or index pages are both purposeful and navigational. Pages immediately subordinate to these often provide navigation to content. All pages provide information; content pages differ from others because their purpose is to transfer the information they contain, rather than to provide definition or direction to that information. Navigation and purpose pages are analogous to tables of content, indexes, and forewords. Content pages are similar to chapters or sections.

The Web sites or pages to which they point can define navigation pages. McDonnell, Koehler, and Carroll (1997) describe four: Jump, gateway, content + 1, and content. Jumppages are eclectic and point to other Web sites. They provide a directory function, much like a bibliographic catalog. Gateways point to cohesive collections of Web pages, often, but not always, on the same site. These are similar to tables of content. Content + 1 pages are found immediately "above" content pages in the directory structure and point to very closely related material. Content pages deliver that material.

A propositus page is that page on which an analysis is centered. The term propositus is borrowed from the genealogical lexicon. It indicates the individual from whom relationships are traced both "forward" and "backward." A Web site propositus differs from an index, home, or subject page in that the former is selected by the viewer/analyst while the Web author creates the latter. It is both possible and valid to select a propositus for analysis that is, or is not, an index or subject page. In this study, Web propositi were selected randomly.

Domain names indicate the "pedigree" of a URL. Domain names consist of two or more URL fragments. These are the portion of the URL between the transfer protocol (e.g., http, ftp) and the directory structure. They consist of two or more URL fragments. The top-level domain (TLD), the right-most tag, may indicate a functional (com, edu, gov, mil, org, or net) or geographic (ISO 3166 two-letter tags)[2] "publisher." The server-level domain name (SLD) indicates the full publisher and server address. TLDs can be divided into two general groups: Functional and geographic. The functional TLDs identify the type of publisher: .com—commercial; .edu—educational; .gov—US government; .mil—U.S. military; .org—non-governmental organization; and .net—network provider. Additional functional TLDs have been proposed, and these and others are likely to be implemented over time.

The geographic TLDs employ the two-letter International Standards Organization Standard 3166 to identify the country or region of publication. A number of ISO 3166 countries also employ 2LD practices to provide a functional indicator. For example, the fragment ac.uk indicates a British academic server, co.jp signifies a commercial Japanese site, and .gob.mx indicates a Mexican government site. There are a number of variations and mixed usages, and these are described in Koehler and Barnett (1998). About 6% of all geographic TLDs are .us with a United States origin.

Finally, while it is generally true that most functional TLDs originate in the United States, it is not always so. Three functional TLDs in the sample are cases of non-U.S. content. There are numerous examples of non-U.S. material carrying functional TLDs. Examples are: www.republicofnamibia.com—Namibia's official home page; www.catmondo.com—a commercial server in Nepal; www.arab.com—a London-based server providing commercial and government links to Arab countries; www.lanka.net—a Sri Lanka Internet service provider's homepage; and www.guyana.com—a quasi-official document pointing to Guyanese materials.

There are also a number of implicit or "inferred" markers. Academic and other servers can sometimes be identified at the 3LD or 2LD where the 2LD practices are not applied because the university or other name is included in the URL. Examples include mcgill.ca—McGill University in Canada; u-bonn.de—Bonn University in Germany; u-nancy.fr—University of Nancy in France; unam.mx—National Autonomous University of Mexico; and conicyt.cl—the National Commission for the Investigation of Science and Technology in Chile.

5.0. Methodology

Two related WWW-based data sets were collected between December 1996 and January 1998 to begin to map Web page and Web site behavior over time. The first captured data at two points on document, page, audio, video, gopher, ftp, E-mail, and structural statistics for Web sites. The second collected weekly data on page size and link changes for Web pages. A number of attributes were examined to assess the growth, change, and death of those Web pages and Web sites. This section describes the selection of the sample and the software tools used to collect the data upon which the conclusions are based.

The World Wide Web and the tools used to explore and exploit it are dynamic and changing. To insure data consistency, once data collection began, new Web pages and Web sites were not added to the collection. Neither were the two primary data capture software tools, FlashSite 1.01 and WebAnalyzer 2.0 changed or upgraded.

[2] The HotBot search engine at http://www.hotbot.com provides a list of ISO 3166 bigraphs.

5.1. Selection of URLs

A random selection of 361 URLs was made in the last 2 weeks of December 1996. The WebCrawler random URL generator was the primary tool used. It was selected after other alleged random URL generators were tested and rejected as non-random, or after it was determined that the generator was no longer available. The WebCrawler random URL generator is either also not truly random in its selection of URLs, or it selects URLs from a URL index that is not representative of the World Wide Web as a whole. URLs on the top-level domain (TLD), .com, are represented in the WebCrawler return sets at a rate far greater than their WWW population. The WebCrawler return sets were augmented through selection of URLs from specific domains using the HotBot Expert search engine. The sample distribution was forced by selecting additional URLs from specific domains generated by the HotBot search engine.

A total of 361 URLs was selected to represent very large populations of unknown sizes. The Web site population was estimated in December 1996 at more than 600,000 servers (NetCraft, 1998) and 9.5 million hosts (NetWizards, 1998). The number of Web documents was estimated at more than 500 million (Koehler, 1996). It is generally recognized that larger samples (s) provide better confidence intervals and, therefore, that they better statistically represent the population (σ) from which they are drawn than do smaller samples (e.g., Hays, 1973, pp. 389–413; Johnson & Joslyn, 1995, pp. 171–195). As a rule of thumb, Hays (1973, pp. 395–396) suggests a minimum sample size of 40 to provide an adequate confidence range for samples representing large populations of unknown size. Samples greater than 100 approach the population. For purposes of this research, samples of more than 100 Web pages and sites were generated because, as was anticipated, a portion of the samples has disappeared over time. Some attrition and technical difficulties were anticipated accessing sites and pages over time, and for that reason also, a larger sample was developed.

Furthermore, the Web consists of many sites and pages. These sites and pages can be divided along several dimensions. These include Web page and site size, domain, and Web object distributions. As a consequence, the dataset can be sub-setted to reflect these and other dimensions. The larger the sample, the larger the sub-samples, and therefore the greater the confidence in conclusions derived from the sub-sample analyses.

Table 1 shows the distribution of URLs, by top-level domain, identified between December 10, 1996 and January 9, 1997, and subsequently collected weekly (for Web pages) over the study period. Both functional and geographic TLDs are represented with samples greater than 100, and several TLDs (com, edu, and European TLDs) exceed sub-sample sizes larger than 40.

To create the Web page sample, the WebCrawler and HotBot selection process produced a set of 361 URLs. These URLs ranged from zero level server-level domain addresses to fourth level. That is, the URL returned by the random search engine process ranged from those with no directory structure (http://aaa.bbb.ccc) to those at the sub-subfile level (http://aaa.bbb.ccc/www/xxx/yyy/zzz.html). The Web page URLs were retained as returned to test the proposition that the further down the directory structure a URL lay, the less stable it was likely to be. That is, not only would it more likely disappear sooner, it would experience greater content and structural change. By retaining URLs at a variety of directory structure levels, it was possible to test the assumption.

TABLE 1. Sample distribution.

TLD type	Total	Percent
Functional		
com	94	26.0
edu	69	19.1
gov	12	3.3
mil	11	3.0
net	32	8.9
org	9	2.5
IP Number	1	0.3
Geographic (ISO 3166)		
Africa	1	0.3
Asia	7	1.9
Europe	90	24.9
Middle East	1	0.3
North America	18	5.0
Pacific	11	3.0
South America	5	1.4
Total	361	100.0

The same sample was retained for the Web site analysis. However, each URL was shaved up the directory structure to its discontinuity point, if indeed one existed. With rare exception, URLs were shaved to the zero or first directory level. No URLs were retained with structure below the second level (for a discussion of the URL shave technique and its search applications, see Koehler, 1997a).

A total sample of 343 Web sites was retained for analysis. Four from the original 361 could not be captured because, by the time the attempt to download was made, they had disappeared. The remaining 13 would not support the download. They would "blow up" even after repeated tries. There are several possible explanations for this failure. First, a number of Web sites block access to all except those on the same SLD or TLD. One .mil page was included in the Web page sample, but the Web site was excluded because only those users on .mil servers could access the site's index or homepage. Other Web sites require users to provide demographic data before they are granted further access. Password requirements also sometimes prevented the software from mapping the site. Access denial was not uniform with all such sites. Thus, the efficacy of the hosts' qualifying and password software probably dictated a proportion of the access failures. Finally, access was denied to two sites for unidentified reasons.

5.2. Measures of Change: Pages

Once the URLs were selected, each URL was entered into FlashSite 1.01. FlashSite is a product of InContext (http://www.incontext.ca). FlashSite performs two primary functions: First, it will download Websites, Web pages, or it will prepare a map (Site Map) diagramming the Web site. Second, it will periodically check the selected downloaded Web site or Web page against the then-current counterpart. FlashSite then prepares a report of the results of the comparison. FlashSite 1.01 permits the user to select updates from immediate to 1 week apart, and performs those updates automatically. For this research, FlashSite 1.01 was programmed to update the Web page database once a week, during the early hours of Friday mornings. A Pentium 90 MHz machine running Windows95 with an ISDN Internet 128 kbps connection was used for data collection. To date, no software or connectivity problems have been encountered with the download.

The FlashSite report is in three parts: The first reports in kilobytes (kb) the size of the current document download. Second, it reports the number of new links to the target Web document. Third, it lists changed items linked to the target document. These three measures can be used to track Web document metamorphosis. The first measure (size in kb) captures changes in target document content, while the other two capture changes in the structure of the propositus.

In addition, FlashSite presents a non-exportable spreadsheet-like presentation of all URLs, including the status of the most recent download attempt. Those status messages include "complete" and "network error." The "network error" message occurs whenever, and for whatever reason, FlashSite is unable to access, download, and assess content and structural changes. These reasons include slow response, no DNS entry (that is, the server is absent), file-not-found (the specific page is gone), and idiopathic causes. All "network error" messaged URLs were resubmitted twice more, if necessary, through FlashSite each week. Those URLs that did not download successfully were copied to a browser and "manually" checked for status. Thus far, FlashSite has been unable to capture and download only one URL (or 0.3%) from the sample determined to be "live" after a browser check. "Comatose" URLs are retained in the FlashSite file and rechecked weekly at the same time as were the others. This was done to determine the "resurrection" rate of the comatose sites. The term "comatose" is chosen rather than "dead" because there can be no absolute certainty that a URL will not at some time resurface.

5.3. Measures of Change: Sites

There are major technical and methodological differences between the Web page and Web site data collections. Web page data were collected weekly, while the Web site data were only collected twice. The first dataset was taken from mid-December 1996 to early February 1997. A second Web site data set was collected in July and August 1997, and the third in December 1997 and January 1998.

First, the Web page data collection and processing required no more than 8 person hours once the technique had been developed and routinized. The download of a single very large Web site sometimes required more than 8 hours to accomplish.

Second, WebAnalyzer does not report structural link changes in Web entities. It does report the number and size of what it labels Web documents, pictures, videos, and audios. It also reports the number of ftp, gopher, and E-mail links. Thus, the WebAnalyzer research was designed to produce indicators of the magnitude of Web site size changes measured in bytes, and Web entity distribution changes within those Web sites measured in number of entities.

5.4. Data Differences

The data reported by FlashSite 1.01 and WebAnalyzer 2.0 are not quite the same and, therefore, not completely comparable. The two software packages collect dissimilar data. Some of those differences are immediately obvious. FlashSite reports the number of structural changes to a page, while WebAnalyzer reports the number of Web entities found in a Web site.

Both also report size in bytes or kilobytes. Care must be taken here, for the two software applications are not measuring the size of the same thing. FlashSite measures the bytes needed to store a given Web entity, including all of its parts. A Web page measured by FlashSite includes the space needed to store the textual document as well as its attached graphics, any audios and videos, as well as the links to subordinate or superordinate pages. WebAnalyzer, on the other hand, measures the size of each of the components.

FlashSite 1.01 could be used to download, store, and map Web sites. However, for a project of this magnitude, that is not feasible. FlashSite stores its data on the same disk where it resides, while WebAnalyzer data can be exported and held in portable storage. The first set of WebAnalyzer data has been archived on sixty-one 100-megabyte Iomega zip disks. To store that same data on the FlashSite resident disk would require the dedication of a hard disk larger than 6.5 gigabytes.

6.0. Discussion

The discussion section provides the results of the year-long longitudinal study of Web site and Web page changes. It is divided into three parts: First, a report of Web site and Web page descriptive statistics; second, an analysis of Web site and Web page permanence; and third, an exploration of Web site and Web page constancy. Each part is further subdivided to provide data on domains, object dominance, and size characteristics.

6.1. Descriptive Statistics: Web Sites

Before exploring the dynamics of Web site demise, movement, and change, it is necessary to describe the gen-

TABLE 2. Web site statistics of central tendency: First, second, and third collection periods.

	First collection N = 344		Second collection N = 295		Third collection N = 257	
Web object	Mean	SD	Mean	SD	Mean	SD
Levels	4.82	4.88	5.22	4.67	5.48	6.68
Text, number	564.31	1,472.38	889.10	1,850.52	1,075.17	2,448.55
Text, bytes	1,360,733	3,336,292	2,273,367	4,222,757	2,426,540	5,511,814
Graphics, number	181.41	314.49	350.56	585.36	506.06	1,325.43
Graphics, bytes	2,174,769	4,733,840	3,040,965	5,652,770	4,273,173	17,554,155
Audios, number	4.83	37.33	6.91	27.35	7.09	30.33
Videos, number	0.47	3.52	1.29	8.67	2.28	18.10
Ftp, number	12.57	89.49	15.04	81.69	13.92	47.76
Gopher, total	6.27	21.92	6.86	23.41	13.42	115.42
Mail, number	61.35	240.19	105.51	305.16	171.86	744.76
Total objects	833.10	1,712.30	1,375.26	2,395.14	1,681.36	3,773.24
Total bytes	3,539,064	6,480,651	5,314,333	64,803,162	6,294,392	21,406,574

eral object characteristics of Web sites, as well as information that can be garnered from an examination of their URL elements. Once that has been accomplished, it may become possible to understand which types of Web sites are more likely to persist or cease to exist, and which are more likely to change than are others. With the exception of Table 2, general descriptive statistics are reported for the first harvest of Web sites, since these are representative of the Web as a whole at the time of collection, while data at subsequent harvests are not.

Web sites demonstrate a great deal of variability, as is shown in Table 2. It presents statistics for the Web site sample for two measures of central tendency: The mean and the standard deviation for 12 variables, each collected or generated for the three time periods. WebAnalyzer provides a report that can include the number and size in bytes of text documents (called "documents"), graphics (called "pictures"), audios, and videos. It also counts the number of ftp, gopher, and mail objects found within the site. Because of their size, audio and video sizes were not collected for these purposes. It should be noted that limited data were collected on audio and video size, and that they are estimated to require, at minimum, at least 12 megabytes each.

The data presented in Table 2 indicate that the sample and, implicitly, the World Wide Web consists of a large number of relatively small Web objects with low object counts, together with a much smaller number of relatively much larger and higher count objects. In other words, the three samples are skewed to the left. The magnitude of each of the reported standard deviations indicates wide variation in object number and "byte-weight." World Wide Web sites, therefore, represent very complex space.

The "total data" reported for the first sample could be used to estimate the size of the WWW at its time of collection in early 1997. The second and third data collections cannot be so used since they are no longer a representative sample of the Web at subsequent time points. It is not representative because it is an aging set that necessarily excludes all Web sites created after January 1997. The foregoing implies that Web sites can be sized and that they range from the very small to the very large. Two measures of Web site size have been explored: Object number and object byte-weight. An optimal general measure of Web site size would combine both of the more specific measures.

6.1.1. Domains and other url markers

URLs carry markers that can provide significant explicit and implicit information that may be used to identify publishers and authors, and to offer a basis for predicting future behavior. These include: Transmission medium (http://, ftp://, and gopher://); domain name fragments; the identification of non-standard ports; the location of the point of discontinuity on the directory structure; and the use of the tilde as a connector for continuity groups to the larger server-level domain. These are explored because not only can they help identify authority, publisher, and quality, but these markers may also offer predictive value both for permanence and continuity.

The top-level (TLD), often the second level (2LD), and sometimes third level (3LD) and subsequent domain names provide the generic identity of the Web site publisher. These TLD and 2LD tags can be used to differentiate among Web sites, and at least four of the major Web search engines support searching by these URL fragments.

Table 3 provides the original distribution of the Web site

TABLE 3. Web site distribution: First collection by publisher type in percent, N = 344.

Domain	TLD only	By TLD and 2LD	TLD, 2LD, and inferred
Commercial	26.5	30.5	30.5
Educational	19.5	24.7	29.9
Governmental	2.9	4.7	4.9
Military	3.2	3.2	3.2
Network	8.7	9.6	9.9
Organizational	2.6	3.5	3.5
Geographic	36.3	23.5	17.7
IP number	0.3	0.3	0.3

TLDs. The second column, labeled TLD only, shows the sample distribution according to the actual URL TLD. The third column transfers from the ISO 3166 geographic category to the functional category indicated on the 2LD. Thus, a URL ending with .gub.uy would be reclassified under government in the second column. The fourth column, labeled TLD, 2LD, and inferred includes, under the functional categories, those URLs which identify their functions in ways other than the 2LD practice. These include univ-lyon1.fr and leidenuniv.nl.

Through application of the 2LD practices and URL reading, it is possible to improve the functional publisher identification significantly, in this case from 63% identified to 82%. On inspection, most of the remaining 18% of the ISO domains are of a commercial nature and could be so classified. They are not, for purposes of this analysis, because the functional domain cannot be directly inferred from the URL.

Three other URL markers were identified as offering possible explanatory or predictive power to understanding Web site behavior. These three are the directory structure depth or location of the point of discontinuity on the SLD, the use of the tilde ("~") to mark the point of discontinuity, and the indication of a non-standard server port to access the Web site. The directory structure depth is determined by counting the number of slashes ("/") which follow the TLD. An example of a Web site collected at the first level is http://aaa.bbb.ccc/xxx. A Web site connected with a tilde takes the form http://aaa.bbb.ccc/~yyy. A Web site accessed through a non-standard port has the appearance http://aaa.bbb.ccc:00. It is also possible that transmission media might have explanatory power, but only one non-http site was collected.

Most (77.9%) of the Web sites included in this study were collected at SLD, or "zero" level. A significant minority (19.8%) was located at the first level, many of these including those attached with tildes. Finally, 2.3% were collected at the second level.

Web sites attached to SLDs with the tilde, or at depths greater than the SLD, tend to be significantly smaller than are other Web sites ($\chi^2 = 21.3$, $p \leq .001$, $df = 5$; $\chi^2 = 37.4, p \leq .000$, $df = 10$). Moreover, tildes are found more often on network (32.4%), educational (23.3%), and to a lesser degree on unreclassified geographic (18.0%), organizational (8.3%), and commercial SLDs (7.6%). No "tilde attachments" were found on government and military SLDs ($\chi^2 = 22.6$, $p \leq .002$, $df = 7$).

Non-standard ports were found only on 5.8% of educational SLDs. However, these are distributed across all size categories without statistical significance.

6.1.2. *Web site size*

As was shown in Table 2, Web sites vary in size, whether measured in objects or bytes. Ordinal size categories were developed: "Smallest" to "biggest" based on the distance in standard deviations from the collection period mean for each variable. To manage a few but very large outliers, the data were normalized by multiplying each value by log10. Z-scores of the log10 values for the total objects and total bytes data were then combined to create ordinal variables.

TABLE 4. Web site size distribution by index, total objects, and total bytes: First harvest.

Objects/bytes	*N*	Min	Max	Mean
Smallest	10			
Objects		2	8	4.4
Bytes		292	1,740	1,013
Smaller	26			
Objects		2	35	10.5
Bytes		2,206	73,771	23,582
Small	84			
Objects		11	376	49.3
Bytes		20,813	1,723,056	277,791
Average	81			
Objects		82	453	205.3
Bytes		227,339	2,078,648	916,115
Big	118			
Objects		124	7,506	1,215.0
Bytes		693,540	46,669,520	6,107,749
Bigger	21			
Objects		1,231	14,019	3,526.9
Bytes		4,317,667	52,137,697	18,698,808

The calculation is biased toward text and graphic content because while the Web object total includes not only test and graphic objects, it also includes audio, video, ftp, gopher, and mail objects. The total byte data include only text and graphic values. The index might also be faulted in that it treats both number of objects and byte-weights equally. It is noteworthy, however, that there is a strong positive correlation (Pearson's $r = .856$, $p \leq .000$) between the two variables total objects and total bytes. One variable can serve as an adequate surrogate for the other. This process resulted in the following two distributions, as is shown in Table 4.

6.1.3. *Web site object distribution*

Web object types can be reduced to two general types: Those that provide access to information files, and those that provide interpersonal contact. There are, of course, overlaps. Text, graphic, audio, video, ftp, gopher, and telnet objects generally provide file access; while E-mail, IRCs, MUDs, and similar objects provide interpersonal access.

Web objects can also be grouped by similarity of function. Because their functions are similar and for purposes of simplification of presentation, audio and video objects are combined as "multimedia objects." Ftp and gopher objects are likewise combined as "file retrieval objects." The definition of each value for each of the Web objects is based on the mean and standard deviations for each object percent of total objects, as shown in Table 5. The distribution of the multimedia, file retrieval, and particularly the E-mail values are highly left skewed, with a concentration of lower values

TABLE 5. Distribution of individual Web objects to all Web objects in percent: First harvest.

Web object type	Mean	SD
Text	56.1	21.8
Graphic	33.8	21.5
Multimedia	0.078	0.42
File retrieval	0.17	0.53
E-mail	0.76	12.6

with a limited number of large outliers. These values were normalized in order to develop a meaningful ordinal scale.

Web sites can be categorized by the dominant object(s) they contain. Dominance is not based on raw percentage. If it were, many sites would be considered text dominant since, on average, more than half of all objects found in those sites are text. Dominance is based on variation from the mean and standard deviation values for each object type. For further elaboration on the methodology, see Koehler (1997c).

In keeping with the tradition of the WWW and the personal computer culture as a whole, the Web object dominant categories were assigned descriptive but humorous titles. Thus, text dominant sites are labeled "wordsworth"; graphic dominant sites, "coffee-table"; multimedia, "mogul"; ftp/gopher, "retriever"; and E-mail, "post office." Web sites that contained no dominant object type are labeled as "average."

The final distributions of Web site types based on the relative number of each of the Web object types for the December 1996 to February 1997 and the July to August 1997 samples are shown in Table 6.

6.2. *Descriptive Statistics: Web Pages*

Web pages, like Web sites, can be either simple or complex constructs. This section provides a number of basic observations resulting from the collection and analysis of the Web page data.

Three factors, Web page size, depth, and language further temper the interpretation of the data presented in this section. These too are limited in their generalization to the general population of Web pages. They are limited because, in some cases, of low sub-sample size. That, in turn, affects the confidence ranges for those statistics.

TABLE 6. Web site types based on Web object dominance: First harvest.

Web site type	N	Percent
Wordsworth (text dominant)	73	21.2
Coffee-table (graphic dominant)	45	13.1
Mogul (multimedia dominant)	21	6.1
Retriever (Ftp/gopher dominant)	60	17.4
Post Office (E-mail dominant)	4	1.2
Average (no dominant Web object)	141	41.0

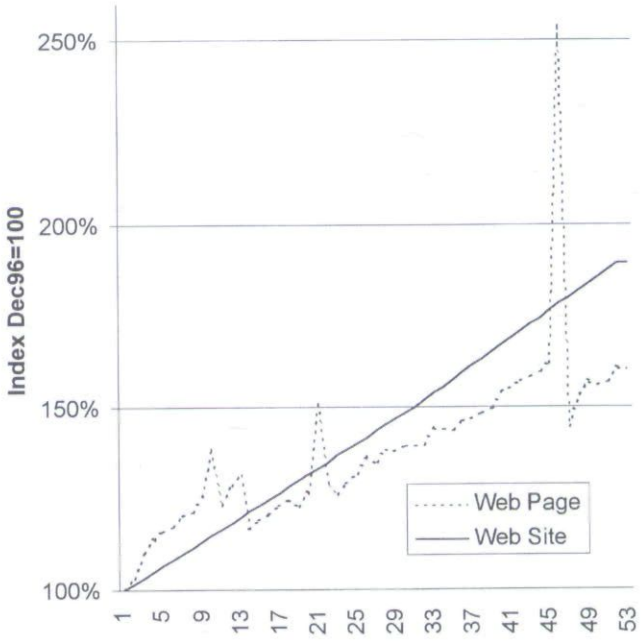

FIG. 1. Average web page and site byte-weight change—December 1996 to January 1998.

6.2.1. *Web page size*

Web page or file size varies dramatically from a zero byte-weight to a maximum of almost three megabytes for this sample. There is no theoretical upward byte-weight boundary. A Web page with zero byte-weight responds to a query but is otherwise empty.

The average size of a Web page at the onset of data collection was approximately 59 kilobytes (kb). Thereafter, the definition of "average size" becomes more complex. The trend over the 53-week collection period was an overall increase in average Web page and site byte-weight or "byte creep." Figure 1 provides byte creep trends for both Web sites and Web pages. The Web page line is indexed to total page byte weight at the first collection, and reports changes weekly. The Web site data are based on three collections, taken in December 1996 to February 1997, July and August 1997, and December 1997 and January 1998. They are the sum of the mean text and graphic byte-weights for each period. Other site values were extrapolated. This represents an annualized rate byte-weight increase for the sample of more than 50% for both pages and sites. Again, these estimates cannot be generalized to the Web as a whole. The sample represents maturing Web page and site groups, and no new documents were added to the samples over the analysis period.

6.2.2. *Web page depth*

The Web page sample is derived from several different directory structure levels on the server-level domain. A quarter was taken from the zero, first, and second levels. Fifteen percent came from the third level. The remaining

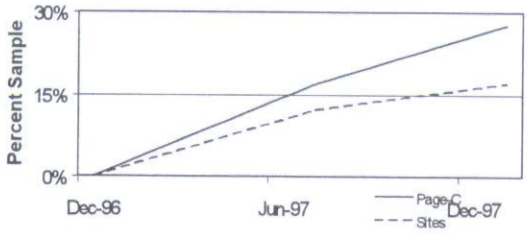

FIG. 2. Web site and page attrition rate—December 1996 to January 1998.

10% were distributed across the fourth to eighth levels. It is unclear whether the distribution of retrieved Web pages is a function of Web site directory structure limits, or if it is a function of an index bias. However, Web sites rarely exceed eight levels, and that is reflected here.

6.3. Permanence

Permanence measures the rate that both Web sites and Web pages continue to respond at the same URL. It does not reflect changes in IP numbers. Nor does it reflect changes to the content of the page or site. It is possible for a Web site or page to change from a large and complex construct to an empty shell, and vice versa without affecting its status as "present." It is also possible for a Web site or page to be intermittently "present" and for its content to remain unchanged when present. Figure 2 plots the attrition rates for the Web page and site samples over the 1-year period. The Web site plot is based on three data points, the page data on 52 weekly points.

The Web page plot necessarily and explicitly includes the Web site data. If Web sites fail, necessarily all Web pages that are a part of that site will also fail. The gap between Web sites and pages represents those Web pages that are removed from extant sites.

6.3.1. Web site permanence

At the end of the second collection, 293 of the original 344 Web sites (85.2%) were still located at the original URL and were collectable. At the same time, 42 (12.2%) had disappeared, while five (1.5%) had moved, and four (1.2%) either denied access or would no longer support harvest by the WebAnalyzer software. By the end of the third data harvest, 74.7% of the Web site sample could be collected at the same URL, while 17.2% were gone. The remaining either forbade access (0.6%), had moved (1.5%), or refused harvest (6.8%). If these trends continue, the half-life of a Web site is estimated to be approximately 2.9 years, while the "accessible" Web site half-life is a somewhat shorter 2.0 years.

Only three data collections were undertaken, therefore the definition of Web site demise for purposes of this analysis is the availability or unavailability of the Web site at the second and third harvests, present and collectable at the first harvest. Harvest was attempted but once each during the second and third collection period. Thus, intermittent Web sites may have been excluded from the analysis. An intermittent Web site is one that fails to respond, and may do so for a variety of reasons. It, however, returns. For example, www.mcgill.ca, the McGill University Web site originating from Montreal, failed to respond the week following the ice storm and subsequent power outages in Quebec in January 1998. It has since returned. The intermittence rate between the second and third collection is 5.2%. That is to say, of the Web sites unavailable for harvest during the second period, 5.2% of the total sample or 35.3% of those unavailable for harvest were available in the third period. This value compares favorably with the Web page intermittence value provided below in Figure 3, suggesting a Web document intermittence "constant."

The analysis of permanence is straightforward. The nominal dependent variable "Availability" is examined for statistical significance against the set of independent variables developed above. These tests are applied against three separate populations: All Web sites present at the first harvest, those Web sites present at both harvests, and those Web sites present at the first harvest but not at the second. Data from both harvests are used as appropriate.

In addition, the two first harvest subsets, those present and those not present at the second harvest, are compared to determine if the two represent similar or different populations. Means and standard deviations for total document, graphic, audio, video, ftp, gopher, and mail objects, as well as text and graphic bytes weights are reported in Table 7. It is noteworthy that in all cases reported in the Table, the means for Web sites unavailable for harvest in July to August 1997 were smaller than those that were available during the same period. Therefore, one rejects the hypothesis, on statistical grounds, that Web sites that persist and those that do not, at least in the short term, represent two different populations based on Web object and byte attributes. One, nevertheless, remains suspicious that smaller object counts and byte-weights may contribute to Web site demise.

There are interesting patterns, also lacking statistical significance, that emerge from an examination of the Web object variables size, density, and object dominance. Larger and denser Web sites are more likely to persist than the smaller and dispersed. Graphic, multimedia, E-mail, and average dominant objects are likely to persist at a slightly

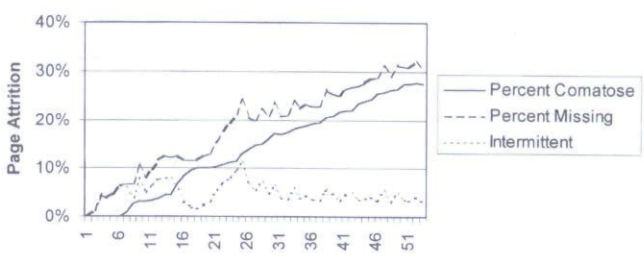

FIG. 3. Web page attrition—December 1996 to January 1998.

TABLE 7. Web site means and SD for available and unavailable sites at second harvest.

	Available Web sites $N = 294$		Unavailable Web sites $N = 50$	
Attributes	Mean	SD	Mean	SD
Text objects	568.6	1,490.7	538.8	1,373.5
Graphic objects	189.5	317.4	130.0	292.7
Audios	5.6	40.3	0.6	2.11
Videos	0.5	3.8	0.3	1.2
Ftp	13.3	96.4	8.4	22.6
Gopher	6.5	23.0	4.7	14.4
E-mail	67.9	258.6	23.1	47.6
Text bytes	1,398,732	3,476,044	1,137,297	2,367,618
Graphic bytes	2,207,039	4,233,552	1,914,527	7,019,269

greater rate, while text and retrieval documents are likely to decline at a slightly greater rate. To stress again, these conclusions lack statistical significance at the 0.10 probability level.

Predictors based on URL derived markers may provide a statistically superior basis for Web site permanence analysis. Three of four are unsatisfactory for general Web site analysis because tildes, ports, and directory depth are not commonly found across all Web site types. However, a substantial number of Web sites where URLs contain tildes, alternate port addresses, or are located at the first directory level failed to thrive ($\chi^2 = 14.1$, $p \leq .000$, $df = 1$; $\chi^2 = 3.5$, $p \leq .062$, $df = 1$; $\chi^2 = 15.9$, $p \leq .000$, $df = 1$, respectively). That may prove useful in anticipating educational, network, and some commercial Web site expirations.

Inferred domain also appears to offer some predictive qualities for Web site permanence. The data provided in Table 8 imply that government and organizational sites tend to be more permanent than other sites. Network and military sites declined at a rate greater than average over the year. Commercial, educational, and the geographic domains, not lending themselves to functional classification on inspection of the URL, declined more or less at the average rate.

6.3.2. Web page permanence

Web pages manifest one of three permanence behaviors. These are: First, that they persisted over the entire sample period without interruption; second, they became "comatose"; or, third, they were intermittent. URLs that "failed to respond" were considered to represent Web pages that were "gone."

The most common "failure to respond" reasons were "No DNS" (no domain name server) and the 400 errors (page not found or access denied). In general, No DNS implies no server. The 400 errors indicate that the page has ceased to exist or is blocked. The No DNS error can occur for a number of reasons. First, the Web site and therefore the Web page had become extinct. Or, the server may be down for any number of reasons ranging from electric failure to political intervention. The 400 errors occur when Web pages but not the site are removed, eliminated, or edited.

As a general rule, most intermittent URLs were the result of DNS problems, while most comatose Web pages were diagnosed by 400 errors. At the end of the sampling period, in January 1998, of the original 361 URL sample, 110 (31%) failed to respond. Of these, 99 (90% of the 110 failing to respond) met the 6-week period "comatose test." Of the 99 meeting the test, 63 (64%) comatose Web pages had been part of Web sites that persisted. This gap between Web site and Web page attrition is illustrated in Figure 2, above.

Of those classed as intermittent failures, in only two instances did Web pages that failed return to persisting Web sites. In all other cases, intermittence was associated with Web site as well as Web page failure. From a practical point of view, the "No DNS" error message may not presage Web document disappearance. The "404 error," on the other hand, almost always does.

The size of the Web may increase over time and the size of Web pages may also, but the number of maturing Web pages decreases over time at a rate of about half a percent a week. Figure 3 charts the no response rate of the sample. The top line is the total non-response rate for each week. The middle line plots the number of Web pages meeting the 6-week period comatose rule. The bottom line indicates the number of intermittent pages for that week. Note that all lines converge on zero in the first week and do not diverge until the sixth. This is an artifact of the initial collection, and the definitions for intermittence and "comatoseness." Figure 3 paints two very interesting pictures. First, the attrition rate of Web pages is virtually linear, and the rate of attrition is about 0.5% per week. At the end of a year, almost 30% of the sample was comatose. It may be that over time, the rate of decline will itself decline. A stable set of mature and persistent Web pages may result. The evidence does not as yet point to that possibility. But identification of any such trend must await further longitudinal analysis.

Second, the number of intermittent pages is relatively constant at about 5% of the sample. The specific pages vary, but the phenomenon does not.

Web pages as well as Web sites occasionally change URLs. Over the study period, 2.2% of the Web pages were

TABLE 8. Web site available by inferred functional domain: First, second, and third harvests in percent.

Domain	First	Second	Third
Commercial	100	81.0	74.3
Educational	100	86.4	73.1
Government	100	100.0	81.3
Military	100	90.9	63.6
Network	100	85.3	68.4
Organization	100	100.0	100.0
Unreclassified ISO	100	85.2	78.9
Average total sample	100	85.5	74.7

TABLE 9. Web page inferred domain by permanence in percent.*

Domain	N	Comatose	Persistent	Intermittent
Commercial	113	19.5	52.2	28.3
Educational	104	11.5	50.0	38.5
Government	18	5.6	66.7	27.8
Military	12	16.7	66.7	16.7
Network	36	27.8	47.2	25.0
Organizational	12	8.3	58.3	33.3
Unclassified ISO 3166	65	20.0	41.5	38.5
Total	360	17.2	50.4	32.4

* Statistic: $\chi^2 = 74.5$, $p \leq .000$, $df = 16$.

forwarded to new or modified URLs. One dropped the alternate port, another added "www" to its address, and the other six moved to entirely new URLs. Each left a forwarding address at the old URL, or automatically and seamlessly transferred the user. These Web pages underwent three different treatments. Of those moved, 25% (two) became comatose shortly after the transfer, 37.5% (three) were modified significantly, and 37.5% (three) underwent no apparent structural or content changes. These Web pages were treated as comatose for purposes of subsequent analysis.

The size of the original Web site is not a particularly good indicator of Web page permanence. There is, however, a tendency for Web pages on smaller Web sites to become comatose more often than those on larger Web sites are. Intermittence is also slightly less likely for Web pages on the larger sites and more likely for the smaller.

Inferred domain is a good predictor of Web page permanence. Several domains are more or less stable than the rest. These are shown in Table 9.

Government and organizational domains are less likely to become comatose than other domains. Government and military domains are also most likely to maintain stability over time, that is, they exhibit less intermittence than other sites as well. Network Web pages are most likely to become comatose than others.

This definition of permanence parallels that given for stability by the Scholarly Societies Project's URL-Stability Index (University of Waterloo, 1997). Government, organizational, educational, and military Web pages are more stable than average while commercial, network, and unclassified geographic sites are less stable.

6.4. *Constancy*

Constancy is a measure of Web site and page content change rather than of their presence or absence. The constancy measures offered in this article do not capture the quality or importance of page or site content changes; they merely capture quantitative indicators.

6.4.1. *Web site constancy*

Almost all Web sites changed in some respect. Based on the sum of total objects and total byte weight, of those Web sites that were available for analysis in the second period, only eight (2.7%) remained completely unchanged over the 192-day period. The other 97.3% each changed in at least one aspect. None had been unchanged in some respect by the end of the third collection.

Even without accounting for the decline in the number of Web sites available for analysis between the first and third collections, both the mean total object count and the total byte-weight more than doubled over the study period, as shown in Table 2. The mean number and mean byte-weight of each of the individual Web objects measured increased without exception. More mature Web sites, it appears, on average, grow larger.

How much larger do those more mature Web sites grow? Total change ranges from quite small to increases and decreases of more than two orders of magnitude. The data are presented in Table 10, and subsequently are based on total objects rather than the sum of total objects and byte-weight. Byte-weight is so much larger than object count that it has the effect of swamping or masking object count variations. Byte-weight is also limited to graphics and text objects. Finally, as already shown, there is a very strong positive correlation between total byte-weight and total objects.

Table 10 paints with very broad strokes a picture of large Web site changes over the two data harvest periods studied. Less than 10% of sites in the sample underwent no change in each period, measured in total objects. Nearly as many sites experienced increases or decreases of total object size by more than 100 times in the first period. This figure does not include those that ceased to exist over the study period, which implies even greater change. Table 10 also demonstrates a second tendency. Overall, Web sites tend to increase in size over time—64.8% increased while 27.8% decreased in the first period, while 77.8% increased in the second and 14.2% decreased.

Table 10 may present interesting data, but it is not subtle. A change of less than one order of magnitude includes all increases or decreases from zero to just less than 10 times the original size. Moreover, the data underlying Table 10 are necessarily not normally distributed and are severely left skewed. These data are percent changes. A change of two negative orders of magnitude is numerically 0.01, where the

TABLE 10. Web site relative changes in orders of magnitude: First and second, and second and third collections.

	Percent	
Order of magnitude	First to second	Second to third
>−2	3.1	0.6
>−1<−2	3.7	6.2
>0<−1	21.0	7.4
No change	7.5	8.0
>0<1	50.2	68.5
>1<2	11.2	6.8
>2	3.4	2.5

TABLE 11. Web site relative change categories: First and second collections.

Category	*N*	Percent
Large decrease	33	11.2
Small decrease	40	13.6
Minute decrease	21	7.1
Nearly no change	43	14.6
Minute increase	58	19.7
Small increase	57	19.3
Large increase	43	14.6

positive is 100. These positive values create statistical outliers.

Table 11 presents a somewhat finer tool. It is based on the normalized distribution of the second sample and first sample total objects ratio, categorized by fractions of the standard deviation. Little or no change is defined as the mean value, plus or minus 0.25 standard deviations. A minute increase/decrease is greater than 0.25 standard deviations to 0.5. A small increase/decrease is from greater than 0.5 to 1.5 standard deviations. And a large increase/decrease is greater than 1.5 standard deviations.

The categories defined in Table 11 are shown because they are employed in the analysis of Web site characteristics as they relate to change.

6.4.2. *Web site constancy and other structural attributes*

Web sites can, as has already been shown, be described according to size, object density, and object dominance, as well as by the URL defined characteristics of domain, port, depth, and tilde.

6.4.2.1. *Size, density, and site constancy.*

The persistent Web site population underwent a shift in size from a "moderate" tendency toward the extremes and the center, over the first collection period to the second. That shift was accompanied by a trend toward extreme size change measured in object numbers. These two phenomena are shown in Table 12.

Web site density is closely related to both variables, since it is a function of both object number and the number of levels within a Web site. Level number and size are highly correlated. Relative change is based on the shift in the number of total Web objects from one period to another. Thus, it is no surprise that the distribution of density and change is similar to size and change. Both first and second period χ^2 values are significant at the 0.05 level.

As indicated by Web site size and density data, Web sites undergo a restructuring as they mature, from a relatively distributed configuration to one that is relatively trimodal. Emphasis shifts to the center and the extremes.

6.4.2.2. *Object dominance and site constancy.*

No major shifts among object dominance types were demonstrated from the first period to the second, with two exceptions. There was again a tendency toward "average." The number of Web sites defined as having an average object distribution increased from 41.4% of the same to 52.5%. Text-dominant objects shifted toward the extremes, large increase and decrease. The balance of the Web sites were redistributed relatively little, reflected by the non-statistically significant statistics each generated (first period: $\chi^2 = 27.1$, $p \leq .617$; second period: $\chi^2 = 39.8$, $p \leq .109$).

TABLE 12. Web site relative changes by first and second period size in row percent.*

Change	Period	Smallest	Smaller	Small	Avg	Big	Bigger
Large decrease							
($N = 33$)	First	6.1	9.1	18.2	15.2	42.4	9.1
	Second	30.3	30.3	30.3	6.1	3.0	
Small decrease							
($N = 40$)	First		2.5	15.0	20.0	50.0	12.5
	Second		10.0	25.0	30.0	35.0	
Minute decrease							
($N = 21$)	First	4.8		23.8	23.8	42.9	4.8
	Second			28.6	23.8	47.6	
Little to no change							
($N = 43$)	First	4.7	14.0	23.3	23.3	25.6	9.3
	Second	4.7	9.3	30.2	27.9	27.9	
Minute increase							
($N = 58$)	First		1.7	12.1	29.3	53.4	3.4
	Second		1.7	12.1	36.2	48.3	1.7
Small increase							
($N = 57$)	First	3.5	7.0	22.8	28.1	31.6	7.0
	Second		5.3	10.5	29.8	45.6	8.8
Large increase							
($N = 43$)	First	7.0	18.6	51.2	16.3	7.0	
	Second			14.0	16.3	67.4	2.3

* Statistics: First period: $\chi^2 = 69.2$, $p \leq .000$, $df = 30$; second period: $\chi^2 = 152.1$, $p \leq .000$, $df = 30$.

TABLE 13. Web site change and domains in percent.*

Change	*N*	Com	Edu	Gov	Mil	Net	Org	Other ISO
Large decrease	33	33.3	3.0			6.1	6.1	51.5
Small decrease	40	30.0	15.0	5.0	5.0	10.0		35.0
Minute decrease	21	14.3	9.5			14.3	9.5	52.4
Little or no change	43	16.3	30.2	7.0	2.3	9.3	2.3	32.6
Minute increase	58	29.3	25.9	5.2	3.4	8.6	1.7	25.9
Small increase	57	21.1	19.3	3.5	5.3	5.3	2.3	42.1
Large increase	43	25.6	18.6		7.0	9.3	2.3	37.2
Total	295	24.7	19.0	3.4	3.7	8.5	3.1	37.6

* Statistics: $\chi^2 = 37.5$, $p \leq .399$, $df = 36$.

6.4.2.3. Domain and Web site constancy. Domain adds little to our understanding of Web site size change. There are, however, several possible generalizations from an analysis of Table 13. The values provided to the right of the relative change variable and below the domain variable are row percentages. Compare these to the total row percentages, the distribution of known and inferred functional domains in the persistent population.

Table 13 suggests four trends. First, organizational, unclassified ISO 3166, and commercial Web sites tended to become smaller. Commercial sites were bimodal. They also tended to increase somewhat. Second, government sites tended to undergo slightly smaller to average changes. Third, educational sites tended toward average to slightly larger changes. And fourth, military sites tended to increase in size.

6.4.2.4. Other URL markers and Web site change. In addition to domain, three other URL markers are explored. These are directory structure depth, alternate ports, and the use of tildes to attach continuities to SLDs. Directory structure depth is a poor indicator of relative change. In part because the port "N" is so small (1.7% of the sample), it contributes little to understanding Web site changes. Web sites carrying tildes will tend toward minute increases and decreases in size, to no changes. Because tildes are closely associated with educational sites, it is no surprise that they predict similar outcomes.

6.4.3. Web page constancy

Web pages undergo significant changes in structure and content over time. The Web page sample was monitored weekly for two types of changes: Content change and structural change. Content change was defined as a change in the byte-weight of Web pages. FlashSite captured the combined text, graphic, audio, and video object byte-weight and reports in kilobytes. Thus, implicitly, any change to Web page content is reflected in the page byte-weight.

Great care must be taken in interpreting and using byte-weight as a surrogate for content. Significant changes in byte-weight need not, and often does not, signal a change in the meaning of any given page. Often it merely signifies the addition of a bandwidth hungry graphic that contributes little but eye appeal to the Web page. At the same time, changes that have little impact on byte-weight can have profound effect on meaning. Change a little punctuation, and meaning can be radically modified. Changes in byte-weight merely signal that changes to content have taken place. They tell us nothing of the importance of those changes.

Structural changes involve modifications to the hypertext links from the propositus Web page. FlashSite reports two types of changes. The first are the additions of new links. The second are modifications in existing links, including their elimination. These too may indicate significant or trivial change in page meaning. There is again no inherent implication to the change count except that change has taken place.

Because it is difficult to ascribe meaning to the amount of structural change found on Web pages, Web page change was captured and analyzed according to whether change occurred or did not. Any amount of content or structural change was considered to have a value of one, no change a value of zero. For the same reasons, the direction of change (positive or negative) at the individual Web page level was likewise not analyzed. The number of Web pages that changed each week could be determined and these were aggregated as a percent of all Web pages.

These aggregations, or what can be termed omega values ($\bar{\omega}$), can be calculated in two ways. The first is to derive an average proportion of changed Web pages over the analysis period by dividing the total raw omega by the number of collection periods. The second method is to divide by the total number of collection periods that the Web page was "present." Both approaches have merit. As has been argued already, it is impossible to ascertain with certainty whether a non-responding Web page is truly and forever gone or whether it is intermittent. If a non-responding intermittent Web page (or, for that matter, a comatose one) is considered to be a part of the sample, then it should be counted both during periods of dormancy and activity. Counting dormant periods for purposes of calculating a Web page's omega has the effect of depressing the omega for each consecutive dormant period and would be captured as either "no change," or zero.

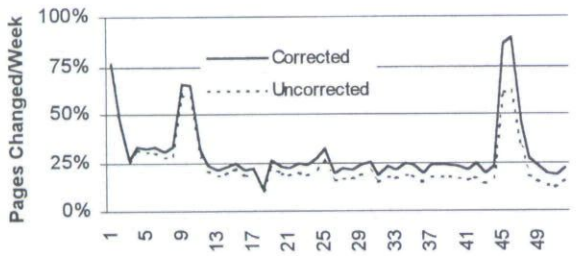

FIG. 4. Web page change rates.

It is equally legitimate to argue that omega values should only be calculated for a Web page when it is active. This is particularly true for those comatose Web pages that have probably dropped off the Web. Calculating omegas based on this approach has the effect of increasing the value. Fortunately, for practical purposes, the differences in values resulting from the two calculations are not that great.

Four sets of omega values were generated. These are $\bar{\omega}_t$, $\bar{\omega}_c$, $\bar{\omega}_n$, and $\bar{\omega}_e$. $\bar{\omega}_t$ is the omega value for all forms of measured content and structural change, $\bar{\omega}_c$ is the omega for content change, $\bar{\omega}_n$ and $\bar{\omega}_e$ are the omegas for new and existing structural change. Each is a measure of change occurring. These values are not additive. Thus, a value of one is assigned total omega, even if at a given time a Web page exhibits content and both types of structural change.

Web pages, like Web sites, manifest greatly different individual patterns of change. For the sample over the 12-month Web page, 3% of the sample returned an overall $\bar{\omega}_t$ of zero. These 11 Web pages had no measured changes. At the same time, 0.8% of the sample changed in at least one respect each time, for an $\bar{\omega}_t$ of one.

The sample average weekly $\bar{\omega}_t$ varied greatly. Figure 4 is a plot of those values over the life of the project. Figure 4 contains two lines. The solid line plots the $\bar{\omega}_t$ as a function of all time periods, whether the Web page responded or not. The dashed line is the same $\bar{\omega}_t$, except that it is calculated

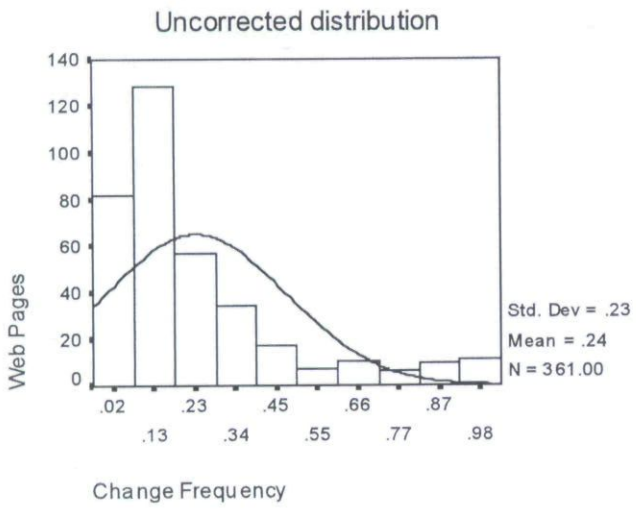

FIG. 5. Web page/change frequency.

TABLE 14. Web page omega values by permanence.

Omega	Total sample	Persistent	Intermittent	Comatose
$\bar{\omega}_t$	0.298	0.271	0.308	0.360
$\bar{\omega}_c$	0.239	0.234	0.221	0.287
$\bar{\omega}_n$	0.074	0.083	0.059	0.075
$\bar{\omega}_e$	0.150	0.176	0.127	0.120

based on the number of periods any given Web page was present. These two lines virtually mirror one another, with the exception that the plot based on periods present is slightly higher than the total periods plot. Because these two sets of values are so highly correlated and so parallel, subsequent analysis is based on a single omega function: Periods present. The plots shown in Figure 4 reflect Web page change of any magnitude and of all types measured. The trends shown in Figure 4 are for Web page change rates of typically about a quarter of "corrected" pages per week. This quiescence is punctuated by brief periods of major change activity. The explanation for these changes in Web page activity is the subject of further research.

At any given time, approximately 25% of Web pages experience change. However, not all Web pages experience change at the same frequency. Figure 5 is a histogram of "uncorrected" Web page changes over the period of study. Very few Web pages remained unchanged over the study period (0.8%) or changed each time (0.6%). As shown in Figure 5, most of the sample changed less than a quarter of the time, and a majority less than 15% of the time. Thus, most Web pages do not change often, but they do change.

Table 14 offers the total sample $\bar{\omega}_t$, $\bar{\omega}_c$, $\bar{\omega}_n$, and $\bar{\omega}_e$ means, and the same values for persistent, intermittent, and comatose Web pages for those periods when the Web pages are present. Comatose and intermittent Web pages do change slightly more overall than the persistent, but not enough to explain the marked early turbulence.

Figure 6 provides plots for the omega values for the three components of change: Content and new and existing structural changes. These data suggest the major turbulent component of Web page change is content change. Both forms of structural change are relatively constant over the study period. Changes to existing structures occur to between 10 and 20% of the sample during each period, while new structural changes occur to between 5 and 10% of the sample for the same period. Another possible explanation

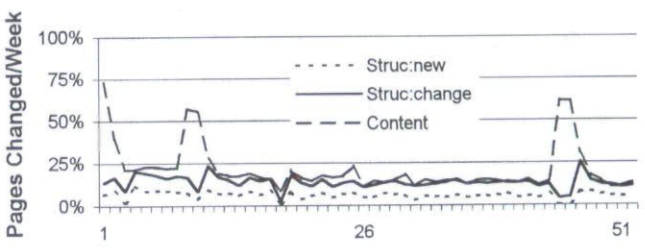

FIG. 6. Weekly web change rates and change types.

TABLE 15. Web page omega and original Web site size.

Omega	Total sample	Smallest	Smaller	Small	Avg	Big	Bigger
$\bar{\omega}_t$	0.298	0.301	0.208	0.308	0.302	0.307	0.286
$\bar{\omega}_c$	0.239	0.253	0.182	0.259	0.226	0.247	0.239
$\bar{\omega}_n$	0.074	0.087	0.044	0.096	0.048	0.085	0.071
$\bar{\omega}_e$	0.150	0.166	0.098	0.156	0.142	0.165	0.128

for the pattern of content change, and one that requires further research to substantiate, is that Web authors tinker with their creations early on. Once satisfied, they are more likely to make fewer and less frequent content changes.

Again, however, while Web page change stability appears to increase over time, it must again be stressed that the degree of change remains high. Even the more mature Web pages experience, on average, a change rate of 20% per week.

6.4.3.1. Original Web site size and Web page change. It was an initial hypothesis that Web site size would have an effect on the frequency of Web page changes: The larger the site, the fewer the changes for any given page. The data presented in Table 15 do not substantiate that hypothesis. The frequency of Web page changes appears to be distributed more or less evenly across all Web site sizes. And indeed there may be unexpected conclusions. Both the smaller and big Web sites manifest Web page changes contrary to the expected.

6.4.3.2. Inferred domain and Web page change. Web pages may vary according to their "publishers," to their different domains. Table 16 offers data according to the inferred domains (TLD, 2LD, and inferred). According to the data in Table 16, Web pages on commercial, military, and network domains experience more change than the sample as a whole. Web pages on educational, organizational, and unclassified geographic domains experience less change. And government domains are about average.

6.4.3.3. Web site object dominance and Web page change. Web site object dominance categories may offer a useful tool in predicting at least some Web page change activity. Omegas for the initial Web site type and the metamorphosis after slightly more than 6 months are offered in Tables 17 and 18. E-mail dominant sites, the data suggest, are more stable than are other object dominant types over time. Multimedia dominant sites, the "moguls," are more prone to early change, but the second period metamorphic forms are no more likely to exhibit greater page change behavior than the others are. The graphic sites, "coffee-tables," are the reverse. Retriever (ftp and gopher) and text dominant sites do not moderate their behavior dramatically, but may become more stable over time.

7.0. Conclusion

The World Wide Web represents a third model, a third information communications vehicle. This article presents data that demonstrate that Web documents differ from oral, unrecorded communications in that they are stored and retrievable, while unrecorded communications by their very nature are not stored and cannot be retrieved. At the same time, Web documents differ from the second form, "traditional," recorded materials in that they are highly ephemeral, do not preserve their intellectual provenance, and are owned and maintained centrally. Web pages, like Web sites, demonstrate a significant variation in permanence and constancy behaviors. Like Web sites, the Web page sizes contract and expand over time. In general, again like Web sites, Web pages on balance are increasing in size—byte creep—thereby contributing to the overall size of the World Wide Web.

That a large proportion ceases to exist over that same period is not in question. Web sites undergo major changes. Two sets of variables were developed to measure and predict Web site permanence and constancy. These variables, derived from the Web site structure and the URLs, are objective metrics which can be captured automatically by commercially available, "low-end" software. These variables, size, density, object dominance, domain, tilde, port, and directory structure depth provided no absolute flags by which to predict with certainty Web site and page behavior. Nevertheless, if interpreted with care, these variables can assist in the understanding of the dynamics of Web document behavior.

TABLE 16. Web page omega and inferred domains.

Omega	Total sample	com	edu	gov	mil	net	org	Uncl geographic
$\bar{\omega}_t$	0.298	0.361	0.255	0.300	0.366	0.327	0.213	0.245
$\bar{\omega}_c$	0.239	0.295	0.186	0.214	0.384	0.300	0.105	0.148
$\bar{\omega}_n$	0.074	0.111	0.048	0.052	0.079	0.105	0.097	0.035
$\bar{\omega}_e$	0.150	0.226	0.105	0.167	0.182	0.170	0.105	0.081

TABLE 17. Web page omega and first Web site object dominance type.

Omega	Total sample	Average	E-mail	Graphic	Multimedia	Retriever	Text
$\bar{\omega}_t$	0.298	0.286	0.111	0.273	0.414	0.301	0.303
$\bar{\omega}_c$	0.239	0.228	0.122	0.222	0.357	0.234	0.247
$\bar{\omega}_n$	0.074	0.074	0.015	0.055	0.052	0.089	0.086
$\bar{\omega}_e$	0.150	0.140	0.051	0.130	0.269	0.152	0.149

Web sites and Web pages undergo significant changes over relatively short periods of time. Three related questions were posed at the beginning of this article:

1. How permanent are Web pages, Web sites, or server-level domains? What is the death rate for each? What is the resurrection rate for each? How often do they move? After a 6-month period, 12.2% of the Web sites and 20.5% of Web pages collected for the study failed to respond when queried, as did 17.7% and 31.8%, respectively, after 1 year. Does this necessarily mean that the half-life of a Web site is about 2.9 years and that of a Web page is about 1.6 years? Continued monitoring is needed to determine whether the attrition rates will continue to decline at established rates or whether an asymptote will be achieved.

Three types of Web page permanence behaviors were identified: Permanence, intermittence, and disappearance. Again, like Web sites, Web pages are often transitory. Web pages, at least for this sample, disappear at a rate of half a percent per week. Both Web pages and Web sites are intermittent. At any given time, approximately 5% of all Web pages are not answering the call, but will return.

2. How constant are Web pages and sites? It was found that more than 97% of Web sites underwent some kind of change over the 6-month period they were followed, after 1 year, more than 99% had changed. In some cases, change was explosive or implosive, in others, quite minor. Web pages likewise undergo changes. Like Web sites, nearly all pages were changed to some degree after 6 months and 1 year (98.3 and 99.1%, respectively). Most change a quarter or less of the time. Very few change every time or not at all.

This study did not document the "importance" or semantics of change. It only took cognizance that some kind of change occurred. Additional research is needed to develop measures of change and the implications of those changes for document meaning.

3. Do different types of Web pages, Web sites, and domains behave differently? Attempts to explain Web page behaviors, like Web site behaviors, meet with mixed results. The original Web site size is not a particularly good predictor of either permanence or constancy. Inferred domain may, on the other hand, contribute to our understanding of Web page behavior. Finally, object dominance may offer greater insights into Web page change behavior and offer slightly more predictive power.

Writing in the late 1930s, H. G. Wells (1938) foresaw the creation of a world brain in a book of the same title. For Wells, the world brain would be a repository of knowledge and knowledge application. If the Internet is truly world brain or its infantile precursor (e.g., Mayer-Kress & Barczys, 1995; Rossman, n.d.), two things can be said for it. World brain has a short memory. And when it does remember, it changes its mind a lot.

This study has explored the "short memory" and "mind changing" of the World Wide Web. A number of indicators are offered. With the exception of one page in the sample, all experienced some degree of content or structural change. A great deal of turbulence was seen throughout the project. World brain may forget or change its mind a lot, but how much and where depends. This study has attempted to begin to chart those patterns.

8.0. Acknowledgments

This article is based on *Web Site and Web Page Permanence and Change: A Longitudinal Study,* MS Thesis, the University of Tennessee, December 1997. The author is indebted to his committee, Carol Tenopir, David Penniman, and George Sinkankas, for their guidance and support. The problem was first identified while developing an Internet cataloging methodology at Information International Associates, Inc. The author also owes a debt of gratitude to its president, Bonnie Carroll, and to his project manager, Janice McDonnell. The anonymous reviewers offered significant suggestions for manuscript improvement and these have been incorporated. The author also thanks them.

TABLE 18. Web page omega and second Web site object dominance type.

Omega	Total sample	Average	E-mail	Graphic	Multimedia	Retriever	Text
$\bar{\omega}_t$	0.298	0.294	0.186	0.358	0.291	0.286	0.259
$\bar{\omega}_c$	0.239	0.238	0.113	0.262	0.256	0.222	0.214
$\bar{\omega}_n$	0.074	0.080	0.036	0.092	0.069	0.065	0.061
$\bar{\omega}_e$	0.150	0.157	0.067	0.199	0.169	0.157	0.122

9.0. References

Chankhunthod, A., Danzig, P., Neerdeals, C., Schwartz, M., & Worrell, K. (1995, November 6). A hierarchical Internet object cache. Available: http://excalabur.usc.edu/cache.html

Chu, H. (1997). Hyperlinks: How well do they represent the intellectual content of digital collections?" In C. Schwartz & M. Rorvig (Eds.), Digital collections: Implications for users, funders, developers, and maintainers, Proceedings of the American Society for Information Science, Vol. 34, Washington, DC, Nov. 1–6, 1997 (pp. 361–369), Medford, NJ: Information Today.

Dunbar, R. (1996). Grooming, gossip and the evolution of language. London: Faber and Faber.

Feldman, S. (1997). "It was here a minute ago!": Archiving the net. *Searcher, 5*(9), 52–64.

Hays, W. (1973). Statistics for the social sciences (2nd ed.). New York: Holt, Rinehart and Winston.

Internet Engineering Task Force. (1997). Uniform resource names (urn). Available: http://www.ietf.org/html/charters/urn-charter.html

Johnson, J., & Joslyn, R. (1995). Political science research methods (3rd ed.). Washington, DC: CQ Press.

Kahle, B. (1997). Preserving the Internet. Scientific American, 276(3), 82–83.

Khan, K., & 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, 176–182.

Koehler, W. (1996). A descriptive analysis of Web document demographics: A first look at language, domain names, and taxonomy in Latin America. In C. Chen (Ed.), Proceedings of the 9th International Conference, New Information Technology, Pretoria, South Africa, November 11–14, 1996 (pp. 159–170), West Newton, MA: MicroUse Information.

Koehler, W. (1997a). An end user's view of mining the Web: Focused and satisfied Internet search and retrieval strategies. Proceedings of the Internet Society Meeting, The Internet: Global Frontiers, Kuala Lumpur. Available: CD-ROM, and http://www.isoc.org/isoc/whatis/conferences/inet/97/proceedings/D3/D3_3.HTM

Koehler, W. (1997b). Document permanence on the WWW. Proceedings of the 1997 Crimea Conference on Libraries and Associations in a Transient World, Sudak, Ukraine.

Koehler, W. (1997c). Web site and Web page permanence and change: A longitudinal study. Unpublished master's, thesis, The University of Tennessee, Knoxville.

Koehler, W. (1998). The librarianship of the Web: Options and opportunities managing transitory materials. In C. Chen (Ed.), Proceedings of the 10th International Conference, New Information Technology, Hanoi, Vietnam, March 24–26, 1998 (pp. 97–106), West Newton, MA: MicroUse Information.

Koehler, W., & Barnett, L. (1998). Domain name searching and World Wide Web search tactics. Searcher, 6(2), 54–60.

Kuhn, T. (1970). The structure of scientific revolutions. Chicago: University of Chicago Press.

Landweber, L. (1997). International connectivity. On the Internet, 3(2), 47.

Lesk, M. (1997) Going digital. Scientific American, 276(3), 58–60.

Lubar, S. (1993). Infoculture: The Smithsonian book of information age inventions. Boston: Houghton Mifflin.

Lynch, C. (1997). Searching the Internet. Scientific American, 276(3), 52–56.

Matrix Information & Directory Services. (1997). Available: http://www.mids.org

Mayer-Kress, G., & Barczys, C. (1995). The global brain as an emergent structure from the worldwide computing network and its implications for modeling. The Information Society, 11(1).

McDonnell, J., Koehler, W., & Carroll, B. (1997). Automating the dynamic development and maintenance of a distributed digital collection. In C. Schwartz & M. Rorvig (Eds.), Digital collections: Implications for users, funders, developers, and maintainers, Proceedings of the American Society for Information Science, Washington, DC, Nov. 1–6, 1997 (pp. 244–259), Medford, NJ: Information Today.

NetCraft. (1998). Available: http://www.netcraft.co.uk/Survey

NetWizards. (1998). Available: http://www.nw.com

OCLC. (1997, February 9). Available: PURL. http://www.purl.org

Price, D. (1963). Little science, big science. New York: Columbia University Press.

Price, D. (1965). Networks of scientific papers. Science, 149(3683), 510–515.

Price, D. (1976). A general theory of bibliometrics and other cumulative advantage processes. Journal of the American Society for Information Science, 27, 292–306.

Rossman, P. (n.d.). World brain. Available: http://www.trib.net/~pressman/wbraib.htm

Rough guide to the Internet. (1977). Available: http://www.asleep.demon-.co.uk/index.htm

Sandar, Z., & Ravitz, J. (Eds.). (1996). Cyberfutures, culture and politics on the information superhighway. New York: New York University Press.

Shuman, B. (1997). Beyond the library of the future: More alternative futures for the public library. Englewood, CO: Unlimited.

Snyder, I. (1997). Hypertext, the electronic labyrinth. New York: New York University Press.

Stefik, M. (Ed.). (1996). Internet dreams: Archetypes, myths, and metaphors. Cambridge: MIT Press.

Tewksbury, R. (1998). Is the Internet heading for a cache crunch? On the Internet, 4(1), 17–22.

University of Waterloo, Scholarly Societies Project. (1997). URL-stability index for the Scholarly Societies Project. Available: http://www.lib.waterloo.ca/society/URL_stability_index.html

Urgo, M. (1996). Analyzing company web sites. InfoManage, the International Management Newsletter for the Information Services Professional, 4(3), 7.

Wells, H. (1938). World brain. Garden City, NY: Doubleday, Doran and Co.

World Wide Web Consortium. (1997, February 3). Names and addresses, URIs, URLs, URNs, URCs. Available: http:www.w3.org/pub/WWW/Addressing/Addressing.html

W3C Architecture Domain. (1997). Learning about URIs. Available: http://www.w3.org/Addressing/Addressing.html

Zakon, R. (1997). Hobbes' internet timeline (Vol. 3.1). Available: http://info.isoc.org/guest/zakon/Internet/History/HIT.html