University of Tennessee, Knoxville

TRACE: Tennessee Research and Creative Exchange

Masters Theses

Graduate School

12-1997

Web site and web page persistence and change : a longitudinal study

Wallace Conrad Koehler

Follow this and additional works at: https://trace.tennessee.edu/utk_gradthes

Recommended Citation

Koehler, Wallace Conrad, "Web site and web page persistence and change : a longitudinal study. " Master's Thesis, University of Tennessee, 1997.
https://trace.tennessee.edu/utk_gradthes/10587

This Thesis is brought to you for free and open access by the Graduate School at TRACE: Tennessee Research and Creative Exchange. It has been accepted for inclusion in Masters Theses by an authorized administrator of TRACE: Tennessee Research and Creative Exchange. For more information, please contact trace@utk.edu.

To the Graduate Council:

I am submitting herewith a thesis written by Wallace Conrad Koehler entitled "Web site and web page persistence and change : a longitudinal study." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Information Sciences.

Carol Tenopir, Major Professor

We have read this thesis and recommend its acceptance:

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a thesis written by Wallace Conrad Koehler, Jr. entitled "Web site and Web page persistence and change: A longitudinal study." I have examined the final copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Information Sciences.

Carol Tenopir, Major Professor

We have read this thesis
and recommend its acceptance:

Accepted for the Council:

Associate Vice Chancellor and
Dean of The Graduate School

WEB SITE AND WEB PAGE PERSISTENCE AND CHANGE: A LONGITUDINAL STUDY

A Thesis
Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Wallace Conrad Koehler, Jr.
December 1997

Copyright © Wallace Conrad Koehler, Jr., 1997

All rights reserved

Abstract

The World Wide Web is, by most accounts, growing and changing rapidly. This research addresses the Web entity issues of life and death, and change over time. This project is concerned with those elements associated with the Web site and Web page URLs and structures that provide insights into Web page and Web site constancy and persistence. It does not address nor analyze the content or meaning of Web pages and Web sites except to the degree that such information can be inferred from the URL.

This analysis is necessarily a longitudinal study. Two data collection periods were established for harvesting the Web site data: December 1996 to February 1997 for the first data capture, and July and August 1997 for the second. Web page data were taken on a weekly basis beginning in early January 1997 and for the purposes of this thesis, ending in late August 1997.

This research also addresses Web entity taxonomies or structures. There has been scant attention paid to Web page or Web site structures. It is suggested that different types of Web entities behave differently. Web pages and Web sites can be distinguished in several ways. This thesis focuses on Web entity attributes that can be determined from an analysis of the URL as well as by measures of Web object types and byte-weight.

It is found that Web sites and Web pages undergo significant changes over time. These changes include the redistribution of object types within Web sites, additions and deletions to text and graphic objects, and additions and deletions of hypertext links to other Web pages. It is also demonstrated that Web site and Web page typologies can help predict constancy and permanence behaviors. The study also suggests that as Web entities mature, their constancy and permanence behaviors moderate somewhat. From one week to the next, approximately twenty percent of Web pages will undergo some degree of change. At the same time, approximately five percent will be intermittently comatose.

Table of Contents

Chapter 1 The Changing Web: A Statement of the Problem 1

Introduction 1

A Short Glossary 3

Web Growth and Change 4

General Web Typologies 7
- Web Content Classification 7
- Web Structure Classification 9

Managing Inconstancy and Impermanence 10

Research Questions 12

Dynamics of WWW Change 13
- Web Pages 13
- Web Sites 14

Implications 16
- Mapping the Web 16
- Diplomatics 17
- Cataloging 18

The Structure of the Study 19

Chapter 2 Methodology 20

Introduction 20

Selection of URLS 20
- Sampling Technique 21
- Random Selection 25
- Two Samples 26
- Measures of Change 27

Data Differences 29

Data Analysis 30

Chapter 3 Web Site Dynamics and Change 31

Introduction 31

Taxonomic Findings 32
- URL Markers 36
- Web Object and Byte-weight Density 39
- Web Object Typology 42

A Web Structure Taxonomy 46

Web Site Longevity 49

Constancy 52
Web Site Constancy and Other Structural Attributes 54
Web Site Persistence and Constancy Conclusions 56

Chapter 4 Web Page Dynamics and Change 58

Introduction 58
Data Collection 58
The Web Page Sample 59
Web Page Size 60
Web Page Depth 60
Web Page Languages 62
Web Page Longevity 63
The coming and going of Web pages over time 64
Change of Address 66
Original Web Site Size and Web Page Persistence 66
Inferred Domain and Web Page Persistence 66
Web Page Constancy 67
Web Page Constancy 67
Original Web Site Size and Web Page Change 73
Inferred Domain and Web Page Change 73
Web Site Object Dominance and Web Page Change 74
Web Page Persistence and Constancy Conclusions 74

Chapter 5 Conclusion 76

Introduction 76
The Web is Different 76
New Approaches 77
This Minor Contribution 78
So What 79

Bibliography 81

Vita 85

List of Tables

Table		Page
2.1	WWW Host, Server, and Document Distributions by Top-Level Domain Name Type – 1996	23
2.2	WebCrawler Random Top-Level Domain Distribution As of 1/3/97	25
2.3	Sample Distribution	26
3.1	Web Site Statistics of Central Tendency, First and Second Samples	32
3.2	Web Site Sample Size Distributions, First and Second Samples	34
3.3	Web Site Size Distribution by Index, Total Objects, and Total Bytes	36
3.4	Web Site Distribution by Publisher Type in Percent, N=344	38
3.5	Inferred Domain Densities in Total Web Objects in Percent	39
3.6	Inferred Domain Densities in Total Byte-weight in Percent	40
3.7	Web Object Factor Analysis Varimax Orthogonal Rotation Solution First Data Collection	42
3.8	Distribution of Individual Web Objects to All Web Objects in Percent Both Samples	43
3.9	Web Site Types Based on Web Object Dominance	45
3.10	Web Site Types Based on Web Object Dominance, Edited	46
3.11	Web Site Means and Std Dev for Available and Unavailable Sites Second Harvest, First Harvest Data	51
3.12	Web Site Domain Distribution by Future Site Availability	51
3.13	Web Site Relative Changes in Orders of Magnitude	52
3.14	Web Site Relative Change Categories	53
3.15	Web Site Relative Changes by First and Second Period Size in Row Percent	54
3.16	Web Site Change and Domains	55
4.1	Web Page Inferred Domain by Persistence in Percent	67
4.2	Web Page Omega Values by Persistence	71
4.3	Web Page Omega and Original Web Site Size	73
4.4	Web Page Omega and Inferred Domains	73
4.5	Web Page Omega and First Web Site Object Dominance Type	74
4.6	Web Page Omega and Second Web Site Object Dominance Type	74

Chapter 1

The Changing Web: A Statement of the Problem

INTRODUCTION

If Vinton Cerf is truly the father of the Internet, then just as surely Vannevar Bush is its grandfather and H.G. Wells its great grandfather. Wells, writing in the late 1930s, 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 (Wells 1938). Bush (1945) laid the foundation for hypertext and gave us "memex," a vehicle for associative memory and the synapses of the world brain. Papa Cerf and many others brought to reality the prescience of Wells and Bush.

If the Internet is truly world brain or its infantile precursor (e.g. Mayer-Kress 1995, Rossman 1997), 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 explores the World Wide Web with a focus on the coming and going of and changes to Web pages and Web sites – their "memory" and "mind changing."

The World Wide Web is, by most accounts, growing and changing rapidly. A great deal has been written on the growth of the WWW, but much less on the ephemeral nature or life and death of Web sites and Web pages, and almost nothing on the metamorphosis, death, and resurrection of Web sites and Web pages experience. This research addresses the latter two issues: life and death, and change over time. This project concerns itself with those elements associated with the Web site and Web page URLs and structures that may be employed to provide insights into Web page and Web site constancy and persistence. It does not address nor analyze the content or meaning of Web pages and Web sites except to the degree that such information can be inferred from the URL and its structure.

This analysis is necessarily a longitudinal study. Two data collection periods were established for harvesting the Web site data: December 1996 to February 1997 for the first data capture, and July and August 1997 for the second. Web page data were taken on a weekly basis beginning in early January 1997 and for the purposes of this thesis, ending in late August 1997.

This research also addresses Web entity taxonomies or structures. There has been scant attention paid to Web page or Web site structures. It is suggested that different types of Web entities behave differently. Web entities can be classified using a variety of markers. One such set of markers consists of the elements or fragments incorporated into URLs. URLs take the general form: *transmission medium://server-level domain name/directory structure.* There are a number of variations. Typically, the transmission medium most often found on the WWW is http, or hypertext transfer protocol. In addition, the Web objects gopher, ftp (file transfer protocol), telnet, and smtp (simple mail transfer protocol) are also seen, as are videos and audios. The server-level domain (SLD) name provides the Internet address for the Web document. The SLD contains at least two fragments, separated by "." or dots. The fragment on the right is the Top-Level Domain (TLD). The TLD indicates in general terms the type of "publisher" of the page. These include the functional TLDs .com, .edu, .gov, .mil, .net, .org, and the newly proposed TLDs. Geographic TLDs indicate the country of origin (usually) using the two-letter ISO 3166 standard. Many, but not all domain name registrars in countries using geographic TLDs support or require the use of functional second-level (2LD) domain names (for a discussion of these practices, see Koehler and Barnett 1998). For example, "co.nz" indicates a commercial publisher in New Zealand. Finally, a small percent of URLs indicate a non-standard port, for example: aaa.bbb:00.

The directory structure of the URL can also provide useful information. One study has found a semiotic pattern of word usage indicating the content of various files on functional commercial URLs (Urgo 1996). Other directory structures can be analyzed as well. A number of Web documents are attached to larger Web sites with tildes. These take the form: aaa.bbb/~xxx/.

The directory/file/subfile structure may provide an indicator of longevity and constancy for both Web pages and Web sites. This structure includes the directory addressing information that follows the

first slash after the top-level domain name. These structures indicate the placement of the computer files on the host computer, a placement controlled by the Web site author. As a general rule, the first directory structure elements are introductory and navigation material. The information they provide concerns descriptions of the site and movement within it, where the content lies. The content pages tend to be further down the structure and provide access to the information the Web site was created to impart. Content pages may be more constant over time, that is the content page is changed less than introductory and navigation material, but they are probably less permanent than the others. Web sites may persist over time, but their content will change. As their content changes, the pages containing the dated material will be eliminated. Web sites are also sometimes edited or rearranged. The index or homepage URL may remain the same (URL permanence), but subordinate pages may be renamed or moved (McDonnell, Koehler, and Carroll 1997).

It is also possible to analyze the structure and size of Web sites and Web pages. Web sites and pages consist of a number of different elements: text documents, graphics, audios, videos, gophers, ftp, and mail objects. Sites and pages are also imbedded with hypertext links from the page to other pages both on and off the SLD. Likewise, other pages point to or are linked to the site or page in question. The number, ratio, size, and distribution of various Web objects can be measured using commercially available off-the-shelf software applications (COTS). Changes in the number and mix of Web objects on a Web site or Web page as well as in the structure of hypertext links from the propositus may be important indicators of content change or potential page and site instability.

A Short Glossary

A glossary is needed to specifically define terms. Following the lead of Lewis Carol's cheshire cat, I choose to define the following in the following way:

Directory structure: URLs consist of two parts, the server-level domain address and implicitly or explicitly the directory structure. The directory structure identifies the location of any specific file or subfile within the host computer. The full URL provides the specific address of any given Web page.

Domain: Domains are Web entity addresses that point to hosts. Domains are parsed from the implicit root domain (e.g. InterNic), the top-level domain (TLD), the second- and subsequent-level domains (2LD, 3LD, etc.), to the server-level domain (SLD). The server-level domain is the full address for the host. It need not, however, point to a single computer. It may provide seamless access to multiple computers/hosts with identical content; that is, to mirrored sites.

Propositus. The term "propositus" is borrowed from the genealogical lexicon. It defines the page chosen by the analyst as the center or target. It may vary within any given site. There are a number of Internet terms (e.g. root document, homepage, and index) that have similar meaning. These terms define a specific page at or near the "top" of a Web site. The propositus page, chosen by the analyst, can be located anywhere on the Web site.

Web page: a Web page is a Web resident document which, once accessed, can be scrolled through on a computer screen without resort to use of hypertext navigation.

Web site: A collection of one or more Web pages having a common focus and a construct continuity. As such, any given Web page part of the Web site need not be stored on the same server or host. In addition, there may be more than one Web site resident on any given server-level domain.

World Wide Web: The World Wide Web, Web, or WWW is one of several information, creation, storage, and transmission services provided by the Internet. Strictly speaking, the WWW is a medium that provides Internet access to hypertext using a specific Internet protocol, the hypertext transfer protocol (http). The accessed hypertext is stored on host computers. Those hosts are interconnected via the Internet, one of several computer networks that communicate with one another using specific communications protocols. The term has popularly come to encompass associated hardware (host computers, transmission media, etc.), software (servers, browsers, applications, etc.), as well as the content or meaning of the hypertext documents the WWW provides access to. The term is used in the broader sense here.

WEB GROWTH AND CHANGE

This study focuses on two aspects of Web site and Web page behavior: the persistence or longevity of Web pages and sites and the constancy or changes to 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?

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; Koehler 1997b). Chankhunthod, et al (1995, http://excalabur.usc.edu/cache-html/subsectionstar3_4_0_1.html) report that the mean lifetime of

all WWW objects was over a three 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 is more useful in managing that transfer than the lifetimes of individual Web objects. The Chankhunthod et al study spanned three months and that is an inadequate period for assessing the usefulness of Web site and page content.

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).

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 (Zakon 1997, Rough Guide to the Internet 1997). Larry Landweber (1997) has until recently published a map of international interconnectivity (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 email, bitnet, and Internet interconnectivity on a country by country basis.

Other WWW measures are increasing. NetWizards (1997) publishes a periodic survey of the number of Internet top-level domains and host 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.

Monk and Claffy (1996), Monk and claffy [*sic*] (1996), and Claffy (1996) have written extensively on Internet backbone statistics, the quality of those statistics, and the need for cooperation in the development, compilation, and dissemination of those statistics. Their data include Internet traffic and methodological issues. They too demonstrate dramatic short-term changes both in the development of infrastructure and of Internet use and demand.

There is a growing library and information sciences literature on electronic or digital libraries, selection, bibliographic control, cataloging, and URL persistence and management. Pattie and Cox (1996), for example, provide a valuable collection of essays on electronic document selection and bibliographic control. Dillon and Jul (1996), for example, point to the "fundamental" characteristics of electronic and traditional documents. There are similarities between the two, but there are also important differences, among these document stability and lifetimes (McDonnell, Koehler, and Carroll 1997).

Morgan (1996), writing in Pattie and Cox, examines the transitory or changing nature of Internet or WWW materials and mechanisms to manage those changes. Morgan also addresses gray literatures, and argues that it is possible to capture the electronic gray literature far more efficiently than in the past. Morgan acknowledges persistence and constancy problems but does not undertake to measure those changes.

Pattie and Cox (1996) and their authors argue vigorously for the application of stringent standards to the capture and cataloging of Internet materials. They point to a general absence of the existence of good metadata on most Internet materials. For example, Web authors often fail to use the simplest metadata html fields, including the title tag and the keyword metatag.

GENERAL WEB TYPOLOGIES

Web sites manifest a number of characteristics which can be assessed through the use of software or which may be deduced from an examination of their URLs. There are two general classes of Web typologies. These are based in the first instance on the content of Web documents and in the second on the structure of those sites and pages. Web content classification is beyond the scope of this thesis. The focus is, instead on structural taxonomies and the predictive power of those taxonomies.

Web Content Classification

There are at least four Web content classification groups. The first group is nominal, and classifies Web content according the presence or absence of specific content. The v-chip for television or the various Internet filters for pornographic material perform this binary function.

The second method is to catalog according to page or site meaning. This is a well-established metadata process and is performed in at least three ways. The first method is hierarchical directories. Yahoo!, for example, provides a hierarchical search directory based on its keyword thesaurus. The second is to apply post coordinate index terms to catalog records of Web documents. The proposals include the Dublin Core Elements List and semantic headers (Desai 1997). OCLC, the US Library of Congress, the European DESIRE/SOSIG project, the Australian National University, and many others are developing indexing schemes.

This coding has been performed both implicitly and explicitly. For example, the number of limited area search engines (LASE) on the WWW continues to grow. A LASE is a search engine with a defined but limited competence. LASEs may be limited to geographic areas, subject matter, publications dates, or any other classification scheme. Web material is accepted by each LASE according to its own criteria. The classics LASE at Indiana University, Argos, indexes those documents nominated by recognized peers, a refereeing function. Others include any document that appears to be relevant to the subject.

Another method is to apply a recognized classification code to Web documents for inclusion in catalog records. The NetFirst database, a part of OCLC's FirstSearch service applies Dewey Decimal codes for Web pages in the database (Koehler and Mincey 1996). Two html metatags also permit a Web site author to provide a precoordinate index and abstract for search engine indexing. The html title can also be used to provide indexable fields.

One example is the Platform for Internet Content Selection (PIC) developed by the World Wide Web Consortium (W3C). PICs are sometimes precoordinate indexing filter codes applied to the Web document by the author. They are more frequently post coordinate codes resident either on a proxy server or on enduser hard disks (Salamonsen and Yeo 1997). PICs therefore can have an almost universal application or a very limited one. They act by permitting or denying access to Web material based on the PIC label and specified search criteria.

The third general classification group is by quality, however defined. Quality may be assessed on a number of factors, for example, authority, timeliness, pertinence, stability, purpose, accessibility, resource level (primary, secondary) and others. This subject has been explored extensively by Tillman (1997) who offers a useful checklist and guidelines for assessing quality.

Finally, it has been proposed that Web documents be classified according to their function; that is, do they provide references to other Web documents that contain the desired information, do they act as tables of content to material elsewhere in the Web document, or do they provide information content. McDonnell, Koehler, and Carroll (1997) have termed these documents "jump," "gateway," and "content"

pages, respectively. Others have created virtual libraries, "my bookmarks," or directories of links to relevant collections of Web documents.

Web Structure Classification

Web sites and pages may also be classified according to structure. Structural classifications may be derived either from the structure of the Web document or from its URL. Web documents vary greatly in their structures. Web documents consist of a variety of Web objects. The objects include text documents, graphics, plug-ins, audio and video objects, email devices, as well as access to other files. Access to other files includes hypertext links to other Web documents on or off the target server. They may also include telnet, ftp, and gopher connections. Web documents may also be presented in alternate formats, for example frames and plain text.

Web sites can be classified according to the presence of absence of specific Web objects. The search engine HotBot, for example, permits its users to specify documents that include specific object types for retrieval. Web sites may also be classified according to the number and mix of Web objects found on the Web site.

Web documents may also be classified according to URL elements. These elements include the transmission medium (http, ftp, gopher, and telnet), the domain name structure, and the directory structure of the URL. The use of these structures as a classification tool and therefore as a search tool has been explored on a limited basis (Koehler 1997a, Koehler and Barnett 1998). Four of the major search engines support searching on URL fragments, at least on a limited basis. These are the expert versions of HotBot, AltaVista, Open Text, and Infoseek Ultra.

Chapter 3 presents data derived from the Web site database utilizing a variety of statistical techniques. These findings are applied to develop a general Web site structural taxonomy based on the number, distribution, and size of Web objects found on Web sites as well as the information derived from an analysis of the Web site URL. Structural taxonomy is used as a basis in understanding the dynamics of Web site change and morbidity.

There are a growing number of solutions proposed to manage catalogs and the standards associated with them. These include the OCLC Dublin Core and Scorpion Projects, the European Community sponsored DESIRE/SOSIG project and the work of the Commission on Preservation and Access. Each is seeking to develop and implement standards for the capture and incorporation of Internet materials into larger library and information science collections. This study can contribute to those processes. It not only begins to outline the constancy and permanence of Web sites and pages, in doing so it explores a body of metadata. The metadata concepts developed in this study are limited to those that can be inferred from the URL or from metrics derived from automated analyses of Web page and Web site object structures and changes.

MANAGING INCONSTANCY AND IMPERMANENCE

Inconstancy and impermanence issues have generated some attention. While statistics for WWW constancy and permanence have rarely been collected, the transitory or non-permanent 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).

URNs have been proposed as a vehicle to identify the identity of an Internet resource rather than its location, the function of URLs. The URN would permit the location of any resource to be moved by resolving the location address through an intermediary. URNs might mimic the ISBN, ISSN, or SICI structures (Daniel 1996, IETF 1997). A unique URN might be assigned to each Web document, but that could also entail assigning a new URN to each iteration of the document.

URCs provide standards for metadata for metadata; that is, how metadata are described on Web sites (Daniel 1995). URCs are analogous to the process of describing the functions, order, and uses of metadata categories described by traditional cataloging standards. URCs describe the range of possibilities

rather specifying specific inputs because the metadata requirements of Web documents vary from document to document.

URIs are to result from a combination of URLs, URNs, and URCs (Daniel, n.d.). Thus, URIs are proposed as an umbrella to incorporate the other URxs in a cohesive scheme and to provide meta-pointers to them (W3C 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 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 Web entity experience. PURLs can also point to archived material.

Perhaps the most ambitious solution to the problem of URL change and impermanence are archives (Feldman 1997), including that 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 in Chapter 3 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 one year. Storage alone would be staggering.

For URxs and PURLs to provide solutions for inconstant or impermanent URLs, some underlying archive of "expired" documents must be maintained. Archives are at best problematic. PURLs do offer a solution to the problem of forwarded URLs, those Web documents that have changed their Web addresses.

The PURL can automatically interpret the old address to the new. But, while URL forwarding is a problem, as is shown in Chapter 4, it is the least of the inconstancy issues. This research contributes to the debate by establishing or suggesting the frequency that archives recapture the Web. It further contributes by beginning to define URL and Web site and Web page attributes that predict the frequency of inconstancy or impermanence. It may be possible to establish schedules for Web document recapture or change assessment based on a statistical sampling of different Web attributes and their predictive contributions to change.

RESEARCH QUESTIONS

There is a wide range of possible descriptive analyses of the WWW that can be undertaken. I suggest that the first order descriptive questions to be addressed are:

1. How stable is the information content of the WWW? How often do Web pages and sites change? What changes?

2. How stable 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?

3. Do different types of Web pages, Web sites, and domains behave differently?

To answer this, a taxonomy must and will be built. The elements of the taxonomy include but are not limited to top-level domain types, size, interactivity, content density (ratio of text, graphics, multimedia), structural density (link structure), and function. Web documents, it has been suggested (McDonnell, Koehler, and Carroll 1997) may perform three information access functions. Two point to content (jump and gateway), while the third, content, presents information in any number of formats.

These questions are explored and results reported. This research does not directly explore the issue of overall Web growth, except to suggest that as a byproduct of the research, "byte creep" is also contributing to WWW growth. On the average, both Web sites and Web pages are increasing in size. These

increases in the size of individual pages, in byte creep, and in Web sites contribute to the growth of the WWW; just as new servers, Web sites, hosts, etc. also contribute to that growth.

DYNAMICS OF WWW CHANGE

To chart WWW change, I have undertaken two parallel research projects. The first follows changes in individual Web pages. The second project focuses on Web sites. Because of the difference in magnitude and structure of Websites and Web pages, two different approaches were required. These approaches are described briefly in this chapter and in greater detail in the methodology chapter.

Web Pages

Changes in the structure and size of Web pages were monitored weekly over an eight-month period using appropriate URL maintainer software. A stratified sample of 360 Web pages was generated in late December 1996. The sample was stratified using the top-level domain (TLD) name distribution of Web hosts and Web pages worldwide. The selection process is described in greater detail in the methodology chapter.

There are several commercial off-the-shelf software packages available to monitor Web page change and mortality. Based on work done elsewhere, I have selected InContext's FlashSite 1.01 as the page monitor. It captures changes in page displacement measured in kilobytes (kb) as well as lists new and changed pages. It reports the status of its searches, providing an indicator of page loss or death. Finally, it can be programmed to run automatically from once a minute to once a week. For further details on FlashSite 1.01, see the InContext homepage at http://www.incontext.ca.

Two types of Web page change were identified: structural and content. Structural changes include changes in the number and type of links from the propositus or target page to other Web pages. Content changes include changes to the content of the page rather than to the structure of the page. Text or graphic additions or deletions, editorial changes, and corrections of typographical errors are examples.

One significant contribution that this research can make to our understanding of Web page dynamics is the potential identification of metric measures of significant change, in effect, how much change "matters?" Again, initial research suggests that the absolute or relative size of Web page structural or content change, measured either in bytes or added and deleted links, may not tell us very much about the import of any change. The addition or deletion of graphics can result in major swings in number of bytes recorded for a page. Number changes on hit counters can result in changed measures. These "large" swings may signify very little. On the other hand, the addition or deletion of critical text, the shifting of punctuation resulting in significant meaning or word changes, may have little impact on byte size.

Web Sites

The second parallel project measures the size and content of Web sites. For purposes of this research, I have defined a Web site as a coherent collection of Web pages that have an identifiable common and related set of themes. Often, coherence can be found at the server-level domain, but it may also occur within the directory structure of the URL. The top-level of the Web site is identified either by a clear break in content or purpose between documents in the directory structure of the URL, referred to here as the point or level of discontinuity. For example, the server-level domain URL may point to an Internet service provider (ISP). Subordinate pages may contain content provided by ISP subscribers, but may have no other relationship to the ISP homepage. These may be identified by the tilde ("~") in the URL, but need not be. This form of discontinuity is most common on .edu, .com, and .net domains but is also found on the .org and ISO 3166 domains.

The one major drawback to this definition is that it varies from those which define Web sites as the entire collection of Web documents resident on any given host or server-level domain. The number of hosts is shown in Table 2.1 of the methodology chapter. Using the continuity/discontinuity definition increases necessarily the number of Web entities to be counted, and/or to be considered. Focusing exclusively on the server-level Web entity misses the variation in content that may be found at the host

level. It is analogous to classifying journals as the targets for bibliographic records, rather than the articles found therein. This research offers an opportunity to estimate the number of server-level domains with multiple Web sites defined as continuities. This work does not count the number of discrete Web sites identified at or further down the directory structure of server-level domains. That can be the subject of further research. It does identify the number of server-level domains with single or multiple sites.

Continuity breaks may also be identified where documents are related in the directory structure of the URL, but neither the subordinate nor the superordinate pages are hypertext linked. This can be the result of Web page designer oversight. But it may also be used as a continuity divider. An examination of super- and subordinate page content is required to establish oversight or purpose. Approximately one percent of the Web site sample exhibits this discontinuity characteristic.

Web sites are collected and archived using InContext's WebAnalyzer 2.0. The software can download links outside the propositus server-level domain to any specified depth. Because of file sizes and equipment limits, I have elected not to download and archive any plug-ins, audios or videos. For the same reasons, I have limited the download of links off the propositus server-level domain to the initial linked page. Downloads are limited to textual and graphic material for the entire Web site, although other Web objects are counted. WebAnalyzer stores each document and graphic. It also prepares a byte-intensive report

The software collects links from the user defined propositus or target document downward in the structure of the URL. This process is both time and resource intensive. Large Web sites may contain several thousand pages, myriad links, and in some cases may constitute more than 20 megabytes of data. A single Web site mapping and download may require more than fours hours using an ISDN Internet connection. The largest downloaded Web site contains over 14,300 pages, the smallest zero pages. In the case of the smallest site, the URL responds, but the page contains no data.

WebAnalyzer provides much useful data: number of documents, graphics, audios, videos, FTP, gopher, and mail links resident on the site. It reports the total size of the first four. It provides an indicator

of the structure of the site by reporting the number of levels created by internal links. It also reports the date any given page was last modified.

IMPLICATIONS

This research has both practical and theoretical implications. Its major contribution is descriptive. The WWW literature is both vast and sparse. Much has been written on the joys and trials of the WWW. And much has been written on the overall growth of the Internet and its WWW. Relatively little has been written on its evolution and life cycles. This effort begins to fill that void.

Mapping the Web

It is a cliché to assert that the World Wide Web has exploded and continues to expand at a rapid rate. Very early and tentative findings of this project suggest that not only is the WWW growing as measured by the number of new sites, new hosts, and new servers. The size of existing Web pages and Web sites is also increasing. This research will contribute to our understanding of WWW growth characteristics as it charts changes and the demise of Web pages and Web sites.

In mapping the Web and the characteristics of Web sites and Web pages, a taxonomy of the Web can be undertaken. Web document types have already been identified, based on function or access to content (Koehler 1996, McDonnell, Koehler, and Carroll 1997). Web taxonomies are much more complex than "jump," "gateway," and "content." They can be classified according to the access to information and the type of information they provide. These pages include free access, qualified access, and limited access. For example, the commercial pornographic pages provide two or more levels before opening to content. The first level qualifies the entrant (an adult, not offended by the content, living where it is legal to view such material) and the second demands a password.

Others restrict access to users either on internal accounts (internal restricted access) or to servers with similar domain signatures (domain restricted access). In the latter case, only users on .mil servers may, for example, access certain Web pages on other .mil TLD servers.

Web sites and their content mix may characterize Web pages. The mix includes text, graphics, audios, videos, FTP, gophers, and e-mail links. Some are primarily text, others are graphics heavy. Those with a high proportion of e-mail and telnet linkages may be said to be "interactive," with FTP access can be said to be "distributive," and those with audios and videos "multi-mediates."

Web sites and pages can be characterized according to size. Both can be measured in bytes. The number of links to and from Web pages can be counted. Not only can links to and from Web sites be counted, but also links within Web sites can be counted.

Their depth, or the number of page levels from the propositus to the bottom may also characterize web sites. WebAnalyzer provides these data. It literally provides a map, presented as concentric rings from the propositus of all Web pages on the same server-domain and their linked relationship to one another. Most Web sites contain no more than eight to ten rings or levels, although the most found at a single site thus far exceeds 178. This "lord of the rings," an English university, proved too large to download -- it is one that "got away." A related characteristic is "ring density." Any number of Web pages may be linked to any given Web page in any given Web site. WebAnalyzer does not report the density for each ring, but an average ring density can be calculated.

Diplomatics

Diplomatics is defined as the science of the analysis of the form of documents. The form and structure of any given document, it is argued, suggests its authenticity, its validity, and the quality of its content (Duranti 1989). A Web site opening with a qualifying page followed by a restricted page will almost probably be a pornographic site. A Web page attached to a Web site with a tilde will probably not

be extensive, and it will probably have a tangential but not direct contextual connection to its superordinate page. If it is connected to an .edu domain, it is very likely an undergraduate student homepage.

I posit that the form and structure of Web documents may be quite informative and may point to authority, source, and quality issues. The diplomatics of Web structures may provide some predictive power to the quality and authority of content.

Cataloging

This research can have practical applications for catalogers. There are a number of major projects addressing methodologies for Web cataloging. Among them are the Dublin Center project at OCLC, NetFirst on FirstSearch, work at the Australian National University, the Stanford Digital Libraries Project, CATRIONA, and work at the US Library of Congress. These projects and other electronic cataloging issues are described in Pattie and Cox (1996), who provide a selection of essays on electronic document selection and bibliographic control. There are a number of smaller projects, including the one with which I am associated at Information International Associates, Inc. These projects are described in McDonnell, Koehler and Carroll (1998).

This research can help answer two critical questions: First, are Web pages or Web sites too transitory either in content or existence to be meaningfully cataloged? Is there a subset of Web pages or Web sites that can be identified according to content, source, or structure that are more permanent or more stable than are others?

Second, we have suggested that the half-life of a Web page is something less than a year. Select "quality" Web sites are, on the other hand, less transitory and disappear at the server-level domain at a rate of about ten-percent per year. The specific half-life varies in part according to the location of the page on the directory hierarchy of the URL: the further down the directory hierarchy, the more transitory the page (McDonnell, Koehler, and Carroll 1997).

Other Web pages and Web sites are intermittent. This fact is certainly of interest to the enduser and may require cataloger attention. Intermittentcy rates for specific Web entities, Web entities by domain type, and the WWW can be established. The duration cycle can also be determined.

THE STRUCTURE OF THE STUDY

This study is divided into five chapters. This first chapter is the introduction, literature review, and statement of the problem. The second chapter describes in detail the methodologies and software and hardware application employed for data collection. Chapters three and four present the finds of the Web site and Web page data analysis. The final chapter provides a discussion and conclusions derived from the analysis of the data.

Chapter 2

Methodology

INTRODUCTION

Two related WWW based data sets were collected between December 1996 and August 1997 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 chapter describes the selection of the sample and the software tools used to collect the data upon which the conclusions are based.

This study seeks to document Web page and Web site change. It is, by necessity, a longitudinal study. 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 software tools, FlashSite 1.01 and WebAnalyzer 2.0 changed or upgraded for analysis of the data capture.

SELECTION OF URLS

A random selection of 360 URLs was made in the last two 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. URoulette had, for example, been recommended as a good URL random generator, but during the period of collection it was non-responsive and may no longer exist. The WebCrawler return sets were augmented through selection of URLs from specific domains using the HotBot Expert search engine.

A total of 360 URLs were selected for three reasons. First, 360 is a good round number and I like a pun. Second, and much more importantly, it has been generally established by statisticians that a minimum sample of 320 offers the same sample confidence range as do larger samples for large universes. A sample greater than 320 was generated because, as was anticipated, a portion of the sample disappeared over time. Third, the universe of possible Web pages was in December 1996 at least 60 million (Koehler 1996) and was probably much larger. The number of Web sites was estimated at the beginning of the study at 600,000. By the close of the study that number had more than doubled (NetCraft 1997).

The universe consists of several sub-universes and, as a consequence, the dataset can be subsetted. Finally, some attrition and technical difficulties were anticipated, and for that reason also, a slightly larger sample was developed.

Sampling Technique

Because one purpose of this data collection is to map general WWW characteristics, one of two sampling and selection approaches could be taken. The first is to accept Web pages in the order the random generator presented them. The second is to stratify the sample according to a generally accepted source, published statistics, or data derived from another source.

If the former strategy had been adopted, any statistics derived from that sample and projected to the universe would have had to have been be weighted to reflect the actual distribution of URLs within the universe of URLs.

Because some of the top-level domains (TLD) are relatively rare, and might not occur in a sample of 360, the decision was made to collect a stratified sample. The sample distribution is based upon the distribution of Web hosts and documents in mid to late 1996.

Host and document distributions are selected for the sample stratification model rather than the server distribution because the collection targets are, in one instance Web pages and are more or less analogous to the default definition engine of a Web document employed by the creators of the HotBot search. The Web site sample is similar to a Web host, and in many cases they are the same. Since the host and document distributions virtually parallel one another, as is shown in Table 2.1, it is also not necessary to augment one set or the other to appropriately stratify it.

Table 2.1 presents 1996 data for measures of the distribution of Web hosts, documents, and servers worldwide. These data were collected and reported by NetWizards (http://www.nw.com/zone/ WWW/dist-bynum.html), the author, and NetCraft (http://www.netcraft.com/Survey/) in July, August, and December, 1996, respectively. The HotBot Expert URL fragment search methodology is discussed in Koehler (1997c). The document data report what HotBot defines and returns in its searches as a "document."

In interpreting these data, it must be remembered that the WWW is dynamic and growing. These data capture points in time, and both totals and percentages are likely to change. It is expected, for example, that the number of non-North American and non-Western European hosts, documents, and servers will increase at a rate greater than those will in more developed or already Internet-penetrated areas.

Second, the number of InterNic controlled top-level domain names will very likely expand in 1998. The International Ad Hoc Committee (IAHC) has proposed to expand the number of generic top-level domain names (gTLD) with the addition of seven new TLDs (IAHC, http://www.iahc.org). These new TLDs will augment and replace the overburdened .com TLD. As of this writing, the United States Departments of Commerce, State, and Justice have initiated inquiries into the domain naming process. The addition of new TLDs will probably effect the distribution shown in Table 2.1. Initially, the new gTLDs would have replaced the .com TLD, but there is a strong possibility that the redistribution of

TABLE 2.1. WWW HOST, SERVER, AND DOCUMENT DISTRIBUTIONS BY TOP-LEVEL DOMAIN 1996

Top-Level Domain Name Type	Host Distribution - July Total	Percent Total	WWW Documents - Aug Total	Percent Total	Server Distribution - Dec Total	Percent Total
"Functional" Names						
Commercial (com)	2,430,954	25.8%	13,693,270	25.7%	401,717	63.1%
Educational (edu)	1,793,491	19.0%	10,116,114	19.0%	12,561	2.0%
Government (gov)	312,330	3.3%	1,396,416	2.6%		
Int'l Gov't Org (igo)	1,557	0.0%	3,022	0.0%		
Network (net)	758,597	8.0%	2,874,784	5.4%	20,467	3.2%
Organization (org)	265,327	2.8%	2,151,396	4.0%	30,026	4.7%
Military (mil)	259,791	2.8%	354,008	0.7%	1,203	0.2%
Functional Subtotal	5,822,047	61.8%	30,589,010	57.4%	465,974	73.2%
ISO 3166 Names						
Africa, North	465	0.0%	1,961	0.0%		
Africa, Sub-Saharan	48,484	0.5%	137,889	0.3%		
America, Caribbean	1,470	0.0%	7,309	0.0%		
America, Central	1,835	0.0%	20,560	0.0%		
America, North	621,211	6.6%	3,032,011	5.7%		
America, South	39,888	0.4%	472,870	0.9%		
Antarctica	7	0.0%		0.0%		
Asia	378,605	4.0%	4,355,337	8.2%		
Europe, Eastern	106,194	1.1%	1,346,043	2.5%		
Europe, Western	2,055,966	21.8%	12,050,731	22.6%		
Middle East	31,694	0.3%	142,798	0.3%		
Pacific Region	365,062	3.9%	1,294,679	2.4%		
Geographic Subtotal	3,650,881	38.7%	22,862,188	42.9%	170,998	26.8%
TOTAL	9,424,447		53,313,309		636,972	

Source: Host data NetWizards, http://www.nw.com/zone/WWW/dist-bynum.html
Document Data: HotBot searches by the author
Server Data: NetCraft, http://www.netcraft.com/Survey/

gTLDs will have more sweeping ramifications. These proposed changes can be followed on the iahc-discuss@iahc.org listserv, archived at http://www.iahc.org/iahc-discuss.

These changes may have significant implications for this research. We can expect some, if not all, .com server-level domains and their associated Web sites to metamorphose quickly from the .com TLDs to the new gTLDs once the system is inaugurated. Others will probably take full advantage of an as yet unspecified transition period. Some server-level domains on other TLDs may follow as well.

These recommendations do not apply to the TLD registration and naming convention based on the International Standards Organization's standard 3166 (ISO 3166). ISO 3166 provides two and three letter and three digit codes for countries and regions (for a list of ISO 3166 codes, see ftp://ftp.tic.com/matrix/countries/iso3166-codes). Internet registrars have adopted the two-letter code for their standard. The two-letter TLD indicates the country of origin for those Web servers, hosts, and documents that bear them. As is shown in Table 2.1, the use for the ISO 3166 standard represents almost 39 percent of hosts, 43 percent of Web documents, and 27 percent of servers.

In general, the functional TLDs represent Web entities originating in the United States, while most ISO 3166 entities originate outside the United States. There are significant exceptions to this "rule." Web entities carrying the TLD ".us, .pr, .vi, and .gu" for United States, Puerto Rico, US Virgin Islands, and Guam, represent almost four percent of all ISO 3166 TLDs. At the same time, there are many TLDs registered with functional .com, .org, and .net domains representing non-US national and regional organizations. Some are official or "semi-official:" see for example, http://www.republicofnamibia.com and http://www.nigeria.com respectively. Others are purely commerical: for example, http://www.arabbiz.com for Arab countries or http://www.catmando.com for Nepal. According to a Matrix Information and Directory Services press release dated June 24, 1996, the number of non-US domains registered on the .com, .org. and .net TLDs are increasing at a rate greater than those located in the United States (MIDS Press Release: Global Domain Names Grow Rapidly Worldwide, June 24, 1996, http://www.mids.org/prodomreg.html).

Random Selection

It was determined that the WebCrawler random URL generator is either not truly random in its selection of URLs or it selects URLs from a URL index which is not representative of the World Wide Web as a whole. URLs on the top-level domain (TLD) .com are represented in the WebCrawler returns sets at a rate far greater than on the WWW. Table 2.2, Random WebCrawler Search, reports the distribution of URLs collected from a series of randomly returned sets of ten URLs.

All URLS were collected from each set of ten returned by WebCrawler until 93 .com records had been harvested, representing 26 percent of the projected 360 total sample. As is shown in Table 2.2, the "quota" of .com hits was filled before half of the projected sample had been collected. Thereafter, non-.com URLs were harvested from subsequent WebCrawler sets while their .com URLs were ignored. The quota for each TLD, once filled, suspended further collection on that domain.

TABLE 2.2. WEBCRAWLER RANDOM TOP-LEVEL DOMAIN DISTRIBUTION AS OF 1/3/97

TLD	Number Collected	Percent
.com	93	56.0%
ISO 3166	38	22.9%
.net	14	8.4%
.edu	12	7.2%
.org	3	1.8%
.mil	3	1.8%
.gov	3	1.8%
TOTAL	166	100%

Once 100 WebCrawler sets of ten had been harvested, representing 1,000 total URLs, or three times the target, specific searches were begun in HotBot Expert on the remaining functional (.mil, .gov, .org) and geographic (all) TLD quota requirements. A total of twenty URLs were harvested from the HotBot searches. To attempt to approach a semi-random collection from HotBot, the number of returns option in the search engine was set to 50 per page. Only one record per page was harvested, and new pages were generated until individual quotas were filled.

Table 2.3 shows the actual distribution of URLs, by top-level domain, collected between December 10, 1996 and January 9, 1997. The sample closely approximates the host and document distributions listed in Table 2.1, above.

TABLE 2.3. 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%

Two Samples

Web page sample. The WebCrawler and HotBot selection process produced a set of 360 URLs. These URLs ranged from zero stage server-level domain addresses to fourth stage. 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). To avoid inadvertent researcher bias, the URL as returned by the random engine, in its original form, was retained for the Web page analysis. Only those URLs that, upon testing in a browser, were no longer viable were

dropped from the analysis. Those dropped URLs were replaced by a new sample. Nineteen or approximately five percent of all tested URLs were non-viable.

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.

Web site sample. 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 360 could not be captured because, by the time the attempt to download was made, they had disappeared. The remaining thirteen 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 idiopathic reasons.

Measures of Change

Page Measures. Once the URLs were selected, each URL was entered into FlashSite 1.01. FlashSite is a product of InContext, a Canadian company (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 one 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 first 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 percent) from the sample determined to be "live" after a browser check. "Dead or dormant" 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.

Once a Web page is determined to have been gone, the URL was shaved from the propositus until the shaved URL responded in a browser. The live shaved URL was entered into FlashSite and monitored as a new propositus. This was done to assess the atrophy rate of the URL directory structure.

Web Site Changes. There are profound 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.

The first difference was collection time. The Web page data collection and processing required no more than eight person hours once the technique had been developed and routinized. The download of a single very large Web site sometimes required more than eight hours to accomplish. That task required a dedicated computer. Frankly, I lacked the computer, technical, financial, and particularly the time resources required to accomplish the Web site download more than twice over the research period. I anticipate, however, repeating the data capture once a year to track changes in the chosen Web sites until such time as they cease to exist.

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.

DATA DIFFERENCES

The data reported by FlashSite 1.01 and WebAnalyzer 2.0 are not 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. Moreover, WebAnalyzer provides data that FlashSite does not.

DATA ANALYSIS

The data were collected in a spreadsheet (Excel) , then imported into a statistical package (SPSS) for analysis. Appropriate monovariate, bivariate and multivariate tests were performed. These include reporting ranges, means, modes, medians, and standard deviations.

Each record in the two datasets was coded according to the observed differences: domains, size, directory structure location, distribution of Web entities, and other attributes. Tests of significant difference (t-tests and chi-square), were performed using these differences as appropriate. The results of these statistical tests are reported in the appropriate chapters.

Chapter 3

Web Site Dynamics and Change

INTRODUCTION

The World Wide Web is a dynamic and volatile medium. This Chapter undertakes to document two elements of that volatility: the life and death of Web sites as well as the changes they undergo. This Chapter provides not only statistics for longevity and change, it also reports a range of statistics for Web site URLs and for Web site objects. These include transmission media, domain name elements, and directory structures. The analysis of Web site change focuses on variations over time in the number of text documents and graphics, as well as the following Web entities: audios, videos, ftp, gopher, and mail objects. Data on the number of hypertext levels are also reported.

The term "Web site" is defined here slightly differently than it is usually. The definition is elaborated more fully in Chapter 1. A Web site is considered a group of one or more Web pages that have a common theme or other dominant co-relationship other than co-location on the same host. Thus, there may be more than one Web site per server-level domain or node. Given that definition, however, more than 80 percent of the study sample occurs at the server-level domain.

This Chapter is concerned with the dynamics of change and persistence of Web sites. Before exploring the dynamics of Web site demise, movement, and change, it is necessary to describe the general 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 to change than are others.

TAXONOMIC FINDINGS

Web sites demonstrate a great deal of variability, as is shown in Table 3.1. It presents statistics for the Web site sample for three measures of central tendency: the mean, the median, and the standard deviation for twelve variables, each collected or generated for the two 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 twelve megabytes each.

TABLE 3.1 WEB SITE STATISTICS OF CENTRAL TENDENCY, FIRST AND SECOND SAMPLES

Web Object	First Collection N=344			Second Collection N=295		
	Mean	Std. Dev.	Median	Mean	Std. Dev.	Median
Levels	4.82	4.88	4	5.22	4.67	5
Text, Number	564.31	1472.38	106	889.10	1850.52	194
Text, Bytes	1360733	3336292	222829	2273367	4222757	543552
Graphics, Number	181.41	314.49	52	350.56	585.36	93
Graphics, Bytes	2174769	4733840	465473	3040965	5652770	761049
Audios, Number	4.83	37.33	1663	6.91	27.35	2038
Videos, Number	0.47	3.52	0	1.29	8.67	0
FTP, Number	12.57	89.49	0	15.04	81.69	0
Gopher, Total	6.27	21.92	0	6.86	23.41	0
Mail, Number	61.35	240.19	4	105.51	305.16	12
Total Objects	833.10	1712.30	217	1375.26	2395.14	414
Total Bytes	3539064	6480651	977203	5314333	64803162	1592852

The measure "level" is derived from the relationship between the site's propositus page and the proximity of other objects to it via hypertext links. WebAnalyzer places a Web object on its assigned level according to its most proximate tie to the propositus. The propositus is considered to be the "zero" level. Web objects found at the first level have at least one direct link between them and the propositus. Web objects found at the second level are not directly linked to the propositus, but are linked through one intermediary object. The software presents the level data in two ways. It provides a visual map of the

levels, which includes icons representing the different types of Web objects. InContext call this map a "Wave Front." The software also provides a list of Web objects, one element of which is the level for each object. The hypertext map differs from other Web mapping software, like CLEARWeb or the now non-supported CyberPilot, which provide maps of the directory structure among Web pages. These latter maps do not chart Web objects other than pages, nor do they provide the icons that InContext products do. An analysis of levels can provide insights into the organizational principles underlying various Web sites. That analysis can also provide an indicator of Web site density, a concept explored below.

Finally, data for "total objects" and "total bytes" are presented in Table 3.1. These were calculated simply by summing all objects and all byte data, respectively.

The data presented in Table 3.1 indicate that the sample and, implicitly, the World Wide Web consists of a large number of relatively smaller Web objects with low object counts, together with a much smaller number of relatively much larger and higher count objects. In other words, the two samples are skewed to the left. This conclusion is derived from an inspection of the data for means and medians. Almost without exception, the median figures are smaller than the mean. 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.

Table 3.2 presents data specific to the two samples: minimum and maximum values found and the total number of Web objects and bytes in the two samples for each category. With great care, these data can be extrapolated to the WWW. It is fair to argue that where a minimum value of zero is reported, that the value represents a minimum for the Web as a whole. The same cannot be said for values other than zero, nor can it be assumed that the maximum values reported for the samples reasonably represent the WWW. For example, the maximum number of levels reported in either sample is 59. The largest I have encountered exceeds 178 levels. The software crashed at that point when the document size exceeded storage.

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 set of data cannot be so used since the second set is no longer

a representative sample of the Web at the time of its collection, the third quarter of 1997. It is not representative because it is an aging set that necessarily excludes all Web sites created after January 1997.

Both the first sample mean data presented in Table 3.1 and the total data in Table 3.2 can be used to estimate the size of the WWW at its time of collection. According to data presented in Table 2.1, the WWW consisted of some 637,000 Web sites or hosts in December 1996. If the average number of text documents per Web site is 564.31 and the average number of graphics is 181.41, and if there were indeed 637,000 Web sites, in December 1996, the WWW contained more than 350 million text documents and 115 million graphics. Similarly, the size of the WWW in bytes can be estimated. The average Web site consisted of approximately 3.5 megabytes excluding audios and videos. That is approximately 2.2 million megabytes for the Web as a whole in December 1996. As of this writing, the most recent NetCraft (1997) estimate of the number of Web hosts in August 1997 is more than 1.26 million, almost double the number at the first of the year.

TABLE 3.2 WEB SITE SAMPLE SIZE DISTRIBUTIONS, FIRST AND SECOND SAMPLES

Web Object	First Collection N=344			Second Collection N=295		
	Minimum	Maximum	Total	Minimum	Maximum	Total
Levels	1	59	1657	0	50	1550
Text, Number	1	13464	194122	1	18253	262284
Text, Bytes	292	45916769	4.68E+08	0	33610940	6.71E+08
Graphics, Number	0	2277	62223	0	3813	103414
Graphics, Bytes	0	46953281	7.46E+08	0	31668348	8.97E+08
Audios, Number	0	640	1663	0	295	2038
Videos, Number	0	60	162	0	122	380
FTP, Number	0	1161	4325	0	1161	4437
Gopher, Total	0	228	2156	0	205	2025
Mail, Number	0	3847	21108	0	3847	31124
Total Objects	2	14019	285754	1	19258	405702
Total Bytes	292	52137697	1.21E+08	0	53107776	1.57E+09

The foregoing implies that Web sites can be sized and that they range from the very small to the very large however one elects to measure them. 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.

Generating z-scores of the $\log_{10}$ values for the total objects and total bytes variables for each case developed a general measure of Web site size. This was done for both samples. The $\log_{10}$ values were employed to modify the highly skewed raw values to a more normal distribution. Z-scores were calculated to provide a standard measure for the variables. The z-scores for each variable were then ranked ordinally. Those values plus or minus one-half standard deviation from the mean were assigned the value "4." Increments of one standard deviation from "4" were assigned 3, 2, and 1 for decreasing sizes, and 5, 6, and 7 for increasing sizes. The values were then summed, then divided by two. Finally, the resulting size indexes were assigned "Average" for an average value of 4. The value 3.5 was collapsed into 3, 2.5 into 2, and 1.5 into 1. Similarly, 4.5 was recoded as 5, 5.5 as 6, and 6.5 as 7. The values 1, 2, and 3 were assigned the names "Smallest," "Smaller, and "Small;" while 5, 6, and 7 were labeled "Big," Bigger," and "Biggest."

The algorithm 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. However, another index might be faulted because it does not treat them equally or because it assigns inappropriate weight to one or the other. It is noteworthy, however, that there is a strong positive correlation (r^2=.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 3.3.

TABLE 3.3 WEB SITE SIZE DISTRIBUTION BY INDEX, TOTAL OBJECTS, AND TOTAL BYTES

Objects/		First Data Collection				Second Data Collection		
Bytes	N	Min	Max	Mean	N	Min	Max	Mean
Smallest	10	2	8	4.4	9	1	3	1.7
		292	1740	1013		0	658	332
Smaller	26	2	35	10.5	22	2	28	9.6
		2206	73771	23582		1341	73771	18592
Small	84	11	376	49.3	58	10	214	54.7
		20813	1723056	277791		15853	1427003	297171
Average	81	82	453	205.3	76	98	807	345.0
		227339	2078648	916115		242621	3649779	1505180
Big	118	124	7506	1215.0	120	182	7490	2423.5
		693540	46669520	6107749		650832	43886555	10402636
Bigger	21	1231	14019	3526.9	7	9061	19258	3403.8
		4317667	52137697	18698808		9925884	53107776	16785612

URL Markers

The construct of URLs carries 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://); top-level, second-level, and sometimes third-level 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; these markers may also offer predictive value both for longevity and continuity.

Publisher. 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. 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 – US 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 commerical 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 six percent 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-US content. There are numerous examples of non-US 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 commerical 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" 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.

Table 3.4 provides the distribution of the Web site 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 Implicit 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 percent identified to 82percent.

TABLE 3.4 WEB SITE DISTRIBUTION BY PUBLISHER TYPE IN PERCENT, N=344

Domain	TLD only	By TLD and 2LD	TLD, 2LD, and Implicit
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

Other URL Markers. 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 percent) of the Web sites included in this study were collected at SLD, or "zero" level. A significant minority (19.8 percent) were located at the first level, many of these including those attached with tildes. Finally, 2.3 percent 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 (χ^2=21.3, p≤. 001; χ^2=37.4, p≤. 000). Moreover, tildes are found more often on network (32.4 percent), educational (23.3 percent), and to a lesser degree on unreclassified geographic (18.0 percent), organizational (8.3 percent), and commercial SLDs (7.6 percent). No "tilde attachments" were found on government and military SLDs (χ^2=22.6, p≤. 002).

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

Web Object and Byte-weight Density

A Web site's density is calculated by dividing the total number of Web objects or the byte-weight of the site by the number of levels counted by WebAnalyzer. As reported in Tables 3.1 and 3.2, the minimum number of levels reported during the first data collection period was one, with a maximum of fifty-nine. These larger level counts are somewhat rare since the mean number of levels was 4.82 and the median four. Tables 3.5 and 3.6 provide density data in total objects and by byte-weight.

TABLE 3.5 INFERRED DOMAIN DENSITIES IN TOTAL WEB OBJECTS IN PERCENT

Domain, TLD, 2LD, and Inferred	Total Object Density Period 1 in Percent				Total Object Density Period 2 in Percent			
	N	Low	Average	High	N	Low	Average	High
Com	105	21.0	66.7	12.4	84	21.4	60.7	17.9
Edu	103	12.6	61.2	26.2	89	9.0	47.2	43.8
Gov	17		82.4	17.6	16		81.3	18.8
Mil	11	18.2	54.5	27.3	11		63.6	36.4
Net	33	27.3	63.6	9.1	28	17.9	60.7	21.4
Org	12	8.3	75.0	16.7	12	25.0	41.7	33.3
ISO remainder	61	14.8	77.0	8.2	51	11.8	72.5	15.7
Total Sample	342	16.3	67.3	16.3	291	13.7	59.1	27.1
Statistics	χ^2=22.48 p≤.069				χ^2= 31.63 p≤.002			

TABLE 3.6 INFERRED DOMAIN DENSITIES IN TOTAL BYTE-WEIGHT IN PERCENT

Domain TLD, 2LD, and Inferred	Total Byte Density Period 1 in Percent				Total Byte Density Period 2 in Percent			
	N	Low	Avg	High	N	Low	Avg	High
Com	105	15.2	72.4	12.4	83	16.9	68.7	14.5
Edu	103	13.6	67.0	19.4	89	10.1	53.9	36.0
Gov	17		70.6	29.4	16	6.3	75.0	18.8
Mil	11	27.3	45.5	27.3	11		54.5	45.5
Net	33	18.2	72.7	9.1	28	17.9	67.9	14.3
Org	12	8.3	58.3	33.3	11		63.6	36.4
ISO	61	19.7	73.8	6.6	51	15.7	70.6	13.7
remainder								
Total Sample	342	15.2	69.7	15.2	289	12.8	64.0	23.2
Statistics	χ^2= 18.70 p≤.177				χ^2= 23.60 p≤.023			

The mean densities of the two time periods were 626.3 megabytes and 159.4 Web objects per level for the first sample and 854.8 megabytes and 231.2 objects per level for the second sample. Over time, each level gained byte-weight and an increased number of objects per level. That is, over time, Web sites that persisted grew more complex.

In order to test the significance of density by document size and by document domain, the Web object and byte-weight density data were normalized to a standard distribution then recoded into three ordinal categories. These three categories were light, average, and high density. Average was considered those values falling into the zone defined as the mean plus or minus one standard deviation. Low-density sites are those falling below one standard deviation from the mean, high density are those in excess of one standard deviation. The mean and standard deviation values for the first time period were employed in calculating both the first and second period values.

As expected, there is a strong positive correlation between site size and site density. Correlation statistics (Kendall's τ_c, a measure of correlation between two ordinal variables with different numbers of categories) indicate Kendall's τ_c values of 15.3 (p≤.000) for the size of the first data collection by byte-weight density in the first period, 16.73 (p≤.000) for second size by second byte-weight, 16.67 (p≤.000) for

total object density by first collection size, and 20.01 ($p \leq .000$) for the second object density in the second period.

The relationship between site density and domain is more complex and subtle. Tables 3.5 and 3.6 present data for total object and byte-weight domain density by inferred domains. Inferred domains are discussed in the section above entitled URL Markers. They combine specific TLDs with functional identifications taken from the 2LD together with other functionally identifiable URL fragments. The Period 1density distributions are necessarily symmetrical because density level was defined based on the standard distribution of the first collection period sample. And for the same reason, all reported samples tend toward the Average Density value. What is of interest is the variation from the total sample density values in each sample for each domain. The data presented in Tables 3.5 and 3.6 offer two sets of conclusions. The first is that there is density variation among the inferred domains. The second is that the density variation within domains and for the sample as a whole tends to increase over time.

If total object and byte densities are indeed surrogates for Web site complexity, the commercial, network, and the unclassified geographic domains tend to be less dense and therefore less complex than the average. At the same time, the educational, governmental, and organizational domains are more dense and perhaps more complex than the average. Military domains are more complicated. In the first period, by both measures, military domains split between more and less density. They became denser in the second period.

With the exception of government byte-weight, all domains and the sample as a whole increased in density over the data collection period. As Web sites age, those that survive tend to grow more dense and perhaps therefore more complex. These findings, it should be noted, cannot be generalized to the Web as a whole, since the WWW at any given time contains both new and aged Web sites. These findings can only be generalized to those Web sites that persist over time.

Web Object Typology

Web sites can be classified according to the distribution of Web objects found in the sites. There are two basic approaches to the development of a Web object typology. The first is to consider the total number of each object type in the Web site mix. This can provide a general map of typologies. Table 3.7 presents the Varimax rotated solution of a factor analysis of the text, graphics, audio, video, ftp, gopher, and mail data resulting from the first data collection from December 1996 to February 1997.

TABLE 3.7 WEB OBJECT FACTOR ANALYSIS VARIMAX ORTHOGONAL ROTATION SOLUTION FIRST DATA COLLECTION

Objects	Factor 1	Factor 2
Text Objects	.003	.734
Graphic Objects	.284	.667
Audio	.964	.006
Video	.962	.005
FTP	.270	.286
Mail	.004	.487
Percent Variance Explained	28.7	25.2

The factor analysis solution shown in Table 3.7 maps two distinct factors or groups of Web site types based on the total number of Web objects by type found in each Web site. The first factor, which explains 28.7 percent of the variance, shows very high positive loadings for the two multimedia data elements collected. The second factor provides high loadings for the more "traditional" Web object types: text and graphic objects. The factor analysis has mapped an "avant garde" factor and a traditional one. The two together explain more than fifty percent of the sample variance.

There is an inherent difficulty with this approach. While it captures the relative variation among Web object types and reflects variations in Web site total object count, it is not sensitive to count variations within each Web object type. That a Web site may contain, in absolute terms, more or less Web objects than another and that the ratio among those objects can be mapped, the analysis misses variations within the distribution of each Web object variable.

The second analysis type is based not on total object count, but on the percent of each object type for each Web site compared to the average number of each Web object type for the sample population. For example, 56.1 percent of all Web objects found on the Web sample in period one were text objects. The second analysis type is concerned with variation from that value. A series of ordinal categories were developed based on the mean average number of each Web object and its standard deviation.

I believe that 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 email, IRCs, MUDs, and similar objects provide interpersonal access. It is important to keep these distinctions in mind when developing a Web object typology.

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 3.8. The distribution of the multimedia, file retrieval, and particularly the email values are highly left skewed, with a concentration of lower values with a limited number of large outliers. These values were normalized in order to develop a meaningful ordinal scale.

TABLE 3.8 DISTRIBUTION OF INDIVIDUAL WEB OBJECTS TO ALL WEB OBJECTS IN PERCENT BOTH SAMPLES

Web Object Type	First Sample		Second Sample	
	Mean	Std Dev	Mean	Std Dev
Text	56.1	21.8	57.1	21.4
Graphic	33.8	21.5	32.1	20.0
Multimedia	0.078	0.42	0.075	0.34
File Retrieval	0.17	0.53	0.07	0.34
Email	0.76	12.6	0.83	14.3

The following ordinal values for each of the Web objects were derived from the normalized distribution of Web objects:

Text:	Low Text Content	0-30 percent total objects
	Average Text Content	>30-70 percent total objects
	High Text Content	>70-99.99 percent total objects
	All Text Content	100 percent total objects
Graphics:	Low Graphic Content	0-20 percent total objects
	Average Graphic Content	>20-60 percent total objects
	High Graphic Content	> 60 percent total objects
Multimedia:	Low Multimedia Content	0-0.05 percent total objects
	Average Multimedia Content	>0.05-2.0 percent total objects
	High Multimedia Content	> 2.1 percent total objects
File Retrieval:	Low File Retrieval Content	0-0.005 percent total objects
	Average File Retrieval Content	>0.005-0.01 percent total objects
	High File Retrieval Content	> 0.01 percent total objects
Email:	No Email	0 percent total objects
	Introvert	>0 to 0.15 percent total objects
	Extrovert	>0.15-50 percent total objects
	Post Office	>50 percent total objects

These ordinal variables can be reclassified into nominal categories that reflect these relative Web object distributions. The first group could consist of those Web sites dominated by a specific object type, those where one object ranks high but all others rank low or average. The second group could consist of those Web sites where two or more object types rank high. And finally, the third group could consist of the remainder, those where all rating are either low or average. There would be five categories in the first group: high text, graphics, multimedia, retrieval, or email. The second group is larger and consists of nineteen theoretical combinations of two or more high object counts. The final group, the average group, would have but one category. The result of this exercise is shown in Table 3.9.

TABLE 3.9 WEB SITE TYPES BASED ON WEB OBJECT DOMINANCE

Dominant Web Object(s)	First Data Collection		Second Data Collection	
	N	Percent	N	Percent
Text	73	21.2	70	23.8
Text/Multimedia	1	0.3		
Text/Multimedia/Retrieve	2	0.6	1	0.3
Graphic	45	13.1	26	8.8
Graphic/Multimedia	3	0.9		
Graphic/Multimedia/Retrieve			3	1.0
Multimedia	10	2.9		
Multimedia/Retrieve	5	1.5	21	7.1
Multimedia/Retrieve/Email			1	0.3
Retrieve	39	11.3	10	3.4
Retrieve/Email	2	0.6	1	0.3
Email	2	0.6	3	1.0
Average	141	41.0	157	53.2

The twenty-five possible and the thirteen actual categories presented in Table 3.9 are both cumbersome to create and manage for this sample because of methodological difficulties. The "N's" or individual counts for several categories are quite small. Variable values with limited counts result in problematic statistical outcomes. For purposes of this analysis, the smaller categories can be collapsed into the larger ones.

The choice for value reductions should not be made arbitrarily. In this instance, smaller categories were melded with larger ones first by restricting the meld to a larger one sharing at least one dominant attribute. For example, the category "Retrieve/Email" would not be collapsed into "Graphic," for example. Rather, the choice would be between "Retrieve" or "Email." Final choices were guided by reference to three sets of factor matrices: the Varimax rotated solutions of Web object ratio data for the first sample (presented in Table 3.8), the first sample results based on those Web sites which "survived" until the second collection, and the second sample solution. Categories were collapsed into other categories that shared similar loadings on the same factors. The category with the higher loading received the meld.

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 email, "post office." Average remained "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 3.10.

TABLE 3.10 WEB SITE TYPES BASED ON WEB OBJECT DOMINANCE, EDITED

Web Site Type	First Data Collection		Second Data Collection	
	N	Percent	N	Percent
Wordsworth (Text Dominant)	73	21.2	71	24.1
Coffee-Table (Graphic Dominant)	45	13.1	26	8.8
Mogul (Multimedia Dominant)	21	6.1	25	8.5
Retriever (Ftp/gopher Dominant)	60	17.4	11	3.7
Post Office (Email Dominant)	4	1.2	5	3.7
Average (No Dominant Web Object)	141	41.0	157	53.2

A WEB STRUCTURE TAXONOMY

The proceeding discussion developed a set of five Web site general descriptive variables. These are size, domain, other URL markers (tilde, port, and depth), density, and Web object dominance. The purpose of this section is to further elaborate the interrelationship among these five variables. This interrelationship is described in general terms; it points to tendencies rather than absolutes. The variables are each listed in the order in which they are developed above. The other variables are then listed under them and the relationship between the two discussed. If the crosstabulation results in a χ^2 statistically significant at or below the .05 probability level, the secondary variable name is italicized.

Size (as measured by number of total objects).

By domains. The smallest Web sites tend to be commercial, educational and network. The smaller tend to be commercial and the unclassified ISO 3166 domains. The small sites tend to be educational and commercial. Average sites are commercial and unclassified. The large sites are educational and commercial. The largest Web sites are educational. Because of their low "N's," government, military, and organizational sites are not good predictors of size.

By Tilde. Web sites attached to the SLD with tildes are significantly smaller than those that are not.

By Port. Web sites accessed through alternate ports are significantly smaller than those that are not.

By Depth. Smaller Web sites tend to be located further down the directory structure than larger ones.

By Density. Larger Web sites are more dense than smaller ones.

By Dominant Object Type. Text and average document types dominate all small Web sites. The smallest Web sites are, in addition, graphic and email dominated. The average size Web sites are dominated by multimedia documents. The large Web sites are text and retriever dominated.

Domain (as measured by TLD, 2LD, and inferred domains)

By Size. Commercial, network, and organizational domains tend to be average to small in size. The educational domains tend to be larger. Government and military domains are average to larger. The unclassified domains tend to be of average size.

By Tilde. Web sites attached to SLDs by tildes are found almost exclusively on educational, unclassified, and network domains. They are found to a limited degree on commercial domains, and rarely elsewhere.

By Port. Alternative port access is unusual. It is found almost exclusively on educational and unclassified domains, and these are limited.

By Depth. Almost all Web site propositusses are located at the SLD. Those found at the first directory structure level are usually found on network domains and to a lesser degree, educational, unclassified, and commercial. There are very few attached at the second level, and most of these are commercial.

By Density. Commercial and network domains tend to have low density. Organizational domains have average density. Government domains have average to high density. Educational and military domains tend to have high density.

By Object Type. Commercial domains tend to be graphic and multimedia dominant. Many are also average. Educational domains tend to be text and retriever dominant. Government domains tend to be text and email dominant. Military domains tend to be retriever or average. Network domains tend be text, graphic, or email dominant. Organizational domains tend to be graphic or retrievers. Finally, the unclassified domains are text and multimedia dominant.

Density (measured as total objects divided by number of levels).

By Size. The less dense tend to be the smaller, and conversely, the more dense tend to be the larger.

By Domain. The low-density Web sites tend to be commercial, military, and network. Average density are government and organizational. The higher density sites tend to be the educational, government, and military domains.

By Tilde. Low-density Web sites tend to carry tildes, the others tend not to.

By Port. There appear to be no trends.

By Depth. The less dense the Web site, the greater the likelihood that the propositus will be located at a greater depth.

By Object Type. Low-density Web sites tend to be text, graphic, or email dominant or average. Average density sites tend to be graphic, multimedia, or retriever dominant. High-density sites tend to be text, retriever, or email dominant.

Object Dominance

By Size. Wordsworths are bimodal. They tend to be among the smallest and the largest sites. Coffee-table sites are moderately small to average. Moguls are more likely to be of average size. Retrievers tend to be average to large. Post offices are either very small or large. The average site tends to be just that.

By Domain. Wordsworths tend to be educational, government, network, or unclassified domains. Coffee-tables tend to be commercial, network, organizational, or unclassified. Moguls are primarily commercial. Retrievers tend to be educational, organizational, or to a lesser degree military sites. The post office N is too small to generalize, but tend to be found on government and network sites. Average sites are fairly nondescript, but tend to be commercial and military.

By Tilde. Moguls and post offices tend not to be tilde attached. All other types have an even distribution of tilde usage.

By Port. Although interpretation should be tempered by the low post office "N," twenty-five of post offices have alternate ports. For all others, the figure is less than five percent.

By Depth. There is a slightly better than average likelihood that wordsworths, retrievers, and post offices will be located at a depth greater than the SLD.

By Density. Wordsworths and post offices are bimodal; they are characterized by both high and low density. Coffee-tables tend to be low to average density. Moguls tend to be of average density. Retrievers tend to be of average to higher density. Finally, the average Web object site tends to be of lower density.

The variables examined above manifest a complex interrelationship and define the wide range of variability found among Web sites. Do these variables contribute to an understanding of Web site longevity and change, can they be used to help predict demise and change over time? The next section addresses the longevity of Web sites, the one that follows after is concerned with change.

WEB SITE LONGEVITY

As is discussed in Chapter 2, a sample of 360 URLs was generated for both the Web site and Web page analysis. For a variety of reasons, more fully elaborated in Chapter 2, the first or baseline Web site sample consists of 344 URLs (95 percent of the full sample.) This sample was taken from mid-December 1996 to early February 1997. A second sample was collected in July and August 1997. The average time between the two collections was 192 days or approximately 6.4 months.

At the end of the second collection, 293 of the original 344 Web sites (85.2 percent) were still located at the same URL and were collectable. At the same time, 42 (12.2 percent) had disappeared, while five (1.5 percent) had moved and four (1.2 percent) either denied access or would not support harvest by the WebAnalyzer software.

Only two 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 harvest, present and collectable at the first harvest. Harvest was attempted but once during the second collection period. Thus intermittent Web sites may have been excluded from the analysis.

The analysis of longevity 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 3.11. T-tests were performed on these variables by their availability, and none were statistically significant. However, it is noteworthy that in all cases reported in Table 3.11, that 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, email, and average dominant objects are likely to persist at a slightly 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.

TABLE 3.11 WEB SITE MEANS AND STD DEV FOR AVAILABLE AND UNAVAILABLE SITES SECOND HARVEST, FIRST HARVEST DATA

Attributes	Available Web Sites N=294		Unavailable Web Sites N=50	
	Mean	Std Dev	Mean	Std Dev
Text Objects	568.6	1490.7	538.8	1373.5
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
Email	67.9	258.6	23.1	47.6
Text Bytes	1398732	3476044	1137297	2367618
Graphic Bytes	2207039	4233552	1941527	7019269

Predictors based on URL derived markers may provide a statistically superior basis for Web site persistence 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 (χ^2=14.1, p≤. 000; χ^2=3.5, p≤. 062; χ^2=15.9, p≤. 000 respectively). That may prove useful in anticipating educational, network, and some commercial Web site expirations.

Inferred Domain also appears to offer some predictive qualities. The data provided in Table 3.12 imply that commercial sites expire at a rate greater than the average, that government, organization, and military sites tend to persist at a rate greater than the average, while the balance tend to persist at the average rate.

TABLE 3.12 WEB SITE DOMAIN DISTRIBUTION BY FUTURE SITE AVAILABILITY

Domain	N	Available	Unavailable
Commercial	105	81.0	19.0
Educational	103	86.4	13.6
Government	17	100.0	0.0
Military	11	90.9	9.1
Network	34	85.3	14.7
Organization	12	100	0.0
Unclassified	61	85.2	14.8
Total	343	85.5	14.5

χ^2=12.9, p≤.075

CONSTANCY

The statistics are very different for the dynamics of change for Web sites. Almost all Webpages 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 percent) remained completely unchanged over the 192 day period. The other 97.3 percent each changed in at least one aspect.

Even without accounting for the decline in the number of Web sites available for analysis between the first and second samples, the total object count and the total byte-weight increased by more than twice and more than ten times, respectively over the study period (see Table 3.2). Indeed, as shown in Table 3.1, 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 distribution is shown in Table 3.13. The data presented in Table 3.13 and subsequently are based on total objects rather than the sum of total objects and byte-weight. Byte-weight is so much larger that 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 3.13 WEB SITE RELATIVE CHANGES IN ORDERS OF MAGNITUDE

Order of Magnitude	N	Percent
> -2	9	3.1
>-1<-2	11	3.7
>0<-1	62	21.0
No Change	22	7.5
>0<1	148	50.2
>1<2	33	11.2
>2	10	3.4

Table 3.13 paints with very broad strokes a picture of large and perhaps very significant Web site changes over the six-month period studied. Less than ten percent of sites in the sample underwent no

change, measured in total objects. At the same time, nearly as many sites experienced increases or decreases of total object size by more than 100 times. This figure does not include those that ceased to exist over the study period, which implies even greater change.

Table 3.13 also demonstrates a second tendency. Overall, Web sites tended to increase in size -- 64.8 percent increased while 27.8 percent decreased.

Table 3.13 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 ten times the original size. Moreover, the data underlying Table 3.13 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 positive is 100. These positive values create statistical outliers.

Table 3.14 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.

TABLE 3.14 WEB SITE RELATIVE CHANGE CATEGORIES

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

Table 3.14 reinforces the findings from Table 3.13. The categories defined in Table 3.14 are shown because they are employed in the analysis of Web site characteristics as they relate to change.

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.

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 3.15.

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.

TABLE 3.15 WEB SITE RELATIVE CHANGES BY FIRST AND SECOND PERIOD SIZE IN ROW PERCENT

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

Statistics First Period χ^2=69.2, p≤.000 Second Period χ^2=152.1, p≤.000

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.

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 percent of the same to 52.5 percent. 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 χ^2=27.1, p≤.617; Second Period χ^2=39.8, p≤.109).

Domain and Web Site Constancy. Domain is not a particularly good predictor of relative Web site change, although several generalizations are possible from an analysis of Table 3.16. 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 3.16 WEB SITE CHANGE AND DOMAINS

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 χ^2=37.5, p≤.399

Table 3.16 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. The use of the word "trend" here borders on the trite. But no more than "trend" should be inferred from these data.

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 (χ^2=18.1, p≤.113). In part because the port "N" is so small (1.7 percent of the sample), it contributes little to understanding Web site changes (χ^2=10.0, p≤.126). Web sites carrying tildes will tend toward minute increases and decreases in size to no changes (χ^2=16.0, p≤.014). Because tildes are closely associated with educational sites, it is no surprise that they predict similar outcomes.

WEB SITE PERSISTENCE AND CONSTANCY CONCLUSIONS

That Web sites undergo significant changes over relatively short periods of time and that a large proportion cease to exist over that same period is not in question. After a six month period, 12.2 percent of the URLs collected for the study failed to respond when queried. Does this necessarily mean that after a year, a quarter of Web sites will disappear and that the half-life of a Web site is two years? It may be that as Web sites mature they become more persistent. These data, however, do not answer these questions. Further longitudinal monitoring of these Web sites is required for an answer. But, a 12.2 percent loss over six months is not insignificant.

Web sites were also found to change significantly. A few (1.5 percent) changed URLs. Almost all (97.3 percent) changed in size. They either increased or decreased, a small proportion imploded or exploded by two or more orders of magnitude.

Two sets of variables were developed over a large part of Chapter 3 to measure and predict Web site persistence 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 clear flags by which to predict with certainty Web site behavior. Nevertheless, if interpreted with care, these variables can assist in the understanding of the dynamics of Web site behavior.

Chapter 4

Web Page Dynamics and Change

INTRODUCTION

Web pages are individual parts of Web sites. And like Web sites, Web pages experience mortality and constancy changes. Part of the change Web pages experience can be attributed to variations in Web site activity, but not all of it can. Chapter 4 explores Web page changes, both within the context of Web site changes, but also outside that context.

The analysis and description of Web page change differ from that for Web sites. There are two major reasons that. First, Web page data were collected on a weekly basis for a total of thirty-four data collection periods rather than two. Second, different software and collecting parameters were employed.

Web sites were defined as a continuity of related Web objects, which include text, graphic, audio, video, ftp, gopher, email, and other objects. Web pages can be defined similarly. There are also differences. A Web page is defined as any Web object than can be accessed without the necessary utilization of hypertext navigation tools to move within the page. A Web page can be scrolled through. A Web page is also a collection of Web objects, usually a text object with imbedded graphics and other object types. Hypertext links to other Web pages are included in this collection of object types.

DATA COLLECTION

For the collection of Web page data, InContext's FlashSite 1.01 was used (for additional information, see the InContext homepage at http://www.incontext.ca). FlashSite provides two basic

functions. It downloads and caches Web sites or designated portions for offline browsing. It also serves as a URL maintainer.

FlashSite's URL maintenance function was adapted for the collection of Web page longevity and constancy data. Once a week between midnight and eight a.m. on Friday mornings FlashSite automatically tested each URL in the dataset and provided a status report for each URL. That status report includes the status of the URL, the size of the Web page in kilobytes, and changes in page links.

The page link status report included two parts: a list of changed links and a list of new links. Both the link status reports and Web page data were collected on a weekly basis.

THE WEB PAGE SAMPLE

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

I begin this Chapter with the same caveat provided in Chapter 2. The sampling technique to identify the URLs collected for this analysis is not completely random. This is in large part a function of the non-random nature of the search engine index from which the sample was drawn. The general search engines have come under some scrutiny and have been criticized not only because their indexes provide skewed coverage across all Web pages and sites, they also provide limited depth penetration within Web sites (for example, see Brake 1997, Koehler 1997c, and Pike 1997).

Three factors, Web page size, depth, and language further temper the interpretation of the data presented in this Chapter. These too create problems of generalization to the general population of Web pages.

Web Page Size

Web page sizes vary 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. As is explored in greater detail below, some Web pages may become "comatose," that is failed to respond for at least six successive weeks, or come and go intermittently. Should these comatose sites be included or excluded from the calculation? The diplomatic solution is to provide both. Figure 4.1 is a plot of the weekly average Web page byte-weight from January 10, 1997 to August 29, 1997. The top figure is the average sample byte-weight excluding all the comatose. The bottom trace includes both comatose and intermittent Web pages, assuming a value of zero, into the calculation. The middle line is the average of the top and bottom plots.

The trend over the thirty-four week collection period was an overall increase in average Web page byte-weight, "byte creep," if you will. Depending on the definition for average selected, this represents an annualized rate of increase in Web page byte-weight for the sample of between twenty and eighty percent. Again, these estimates should not be generalized to the Web as a whole. The sample represents a maturing Web page group and no new Web pages were added to the sample over the analysis period.

Web Page Depth

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

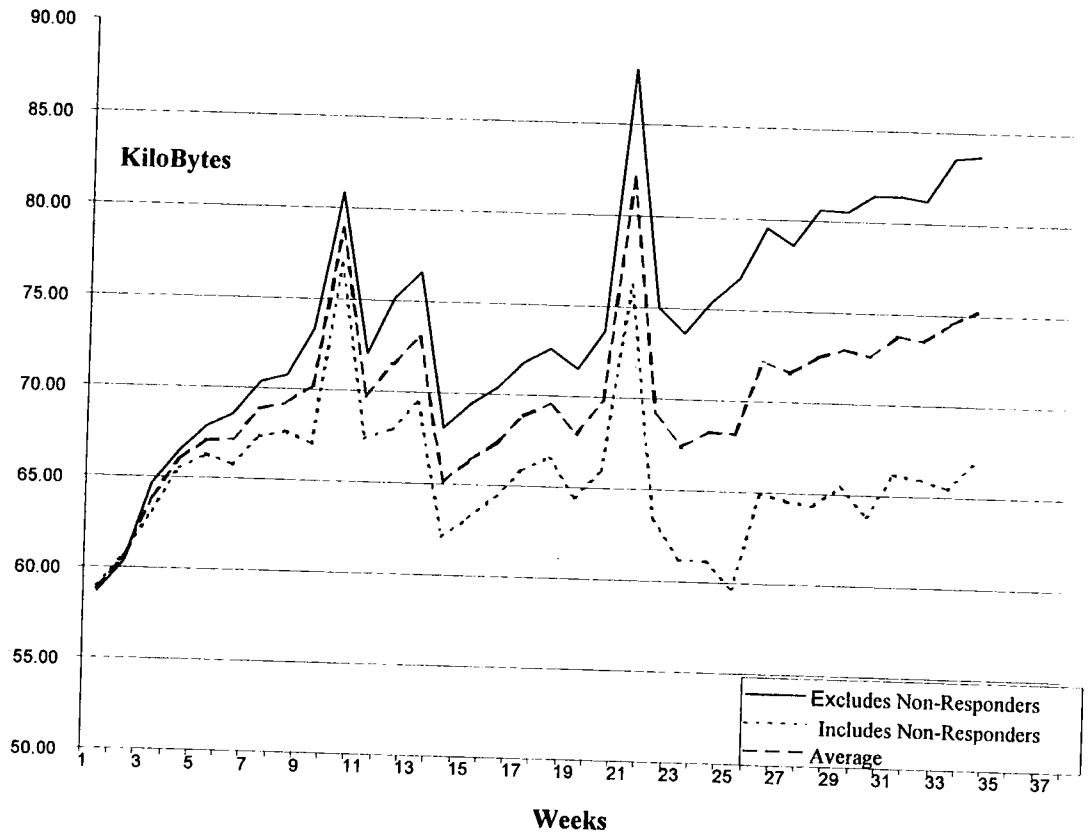

Figure 4.1 Average Web Page Byte-Weight Over Time

Web Page Languages

In general, the Web site and Web page data were not taken from a content analysis of individual pages for presentation and evaluation in this thesis. Language is a special case. Most search engines cannot index and retrieve Web documents written exclusively in pictographic symbols like Chinese, Japanese, or Korean. WebCrawler, the search engine used to generate the sample, is such a search engine. In principle, the search engines can manage non-romance alphabets, but not all do. These limitations further temper the quality of the sample.

Without question, the language found most often on the WWW is English. But it is not the only language. If more than eighty percent of Web pages are published in English, it is because many Web pages are published in countries where English is at least one of the official languages. Not surprisingly, the United States is the source of the majority of Web material, but other English speaking countries are also important contributors. The list includes Australia, Canada, India, Kenya, New Zealand, South Africa, the United Kingdom, as well as many smaller countries (Koehler 1996).

Those countries where English is not the primary language follow two patterns. English, and much less commonly another second language, is often used. But so are the local indigenous and official languages. We found a strong relationship between the sophistication of the domestic telecommunications infrastructure together with the target audience for the Web page and the language used. The lower the sophistication of the telecommunications infrastructure and the more international the target audience, the more likely a Web page would be published in English. Thus, much of the material on African and South Asian servers is in English, while the vast majority of documents on Latin American servers is in Portuguese and Spanish (McDonnell, Koehler, and Carroll 1997).

The languages found on the sample Web pages follow similar patterns. Only one of 229 Web pages on a functional TLD was written in a language other than English. Similarly, with the exception of three of 132, the Latin American, Japanese, and non-English speaking European Web pages were published in local indigenous or official languages.

WEB PAGE LONGEVITY

Web pages manifest one of three longevity behaviors. These are first that they persisted over the entire sample period without interruption; second, they became "comatose;" or, third, they were intermittent.

A comatose Web page is defined as one that failed to respond for six data collection periods or more and which continued to fail to respond through the end of the data collection on August 29, 1997. The term "comatose" was specifically chosen rather than "dead" or "defunct" for these Web pages because it is not possible to state with certainty that any given Web page, once it fails to respond for some specific period, is truly gone. The six period definition was chosen because, with the exception of two Web pages, no Web pages exhibited intermittent behavior over a period greater than six. The two exceptions returned at week twelve and fifteen.

An intermittent Web page is defined as one that failed to respond at least once but returned. Intermittent Web pages also responded within the six periods at the end of the data collection.

Failure to respond was defined as the inability of the software to access the Web page for whatever reason after three tries. The initial try was made during the automatic FlashSite collection on Friday mornings. The second attempt to harvest page data was made Friday afternoons. The third and last attempt was made on Saturdays when the FlashSite data were harvested. If on any of the two subsequent attempts, the absent Web page responded, it was considered "present." This collection methodology eliminated transitory or temporary response failures.

The most common 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 that the page had ceased to exist or was thereafter 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 is 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, on August 29, 1997, of the original 361 URL sample, 83 (23 percent) failed to respond. Of these, 79 (95 percent of the 83 failing to respond) met the six period "comatose test." Of the 79 meeting the test, 51 (64 percent) comatose Web pages had been part of Web sites that persisted. Subsequently, two of those SLDs succumbed. The remaining 36 percent of comatose Web pages "went down" with their SLDs.

Of those classed as intermittent failures, in only two instances did Web pages that failed return to persisting Web sites. In all other case, intermittence was associated with Web site as well as Web page failure. Thus, the failure of the Web page but not the Web site is closely associated with the Web page comatose state. When this condition occurs, the Web page is probably gone. The status of pages when both the page and the SLD fail is more problematic.

The coming and going of Web pages over time

The size of 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 4.2 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 six 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 4.2 paints two very interesting pictures. First, the attrition rate of Web pages is virtually linear, and the rate of attrition is about 0.5 percent per week. At the end of thirty-four weeks, almost twenty percent of the sample had gone. If that projection holds, by the end of a year, the maturing set of pages will decline to sixty percent of the original. By the end of three years, the sample could disappear. It may be that over time, the rate of decline will itself decline. The evidence does not as yet point to that possibility. A stable set of mature and persistent Web pages may result. But identification of any such trend must await further longitudinal analysis.

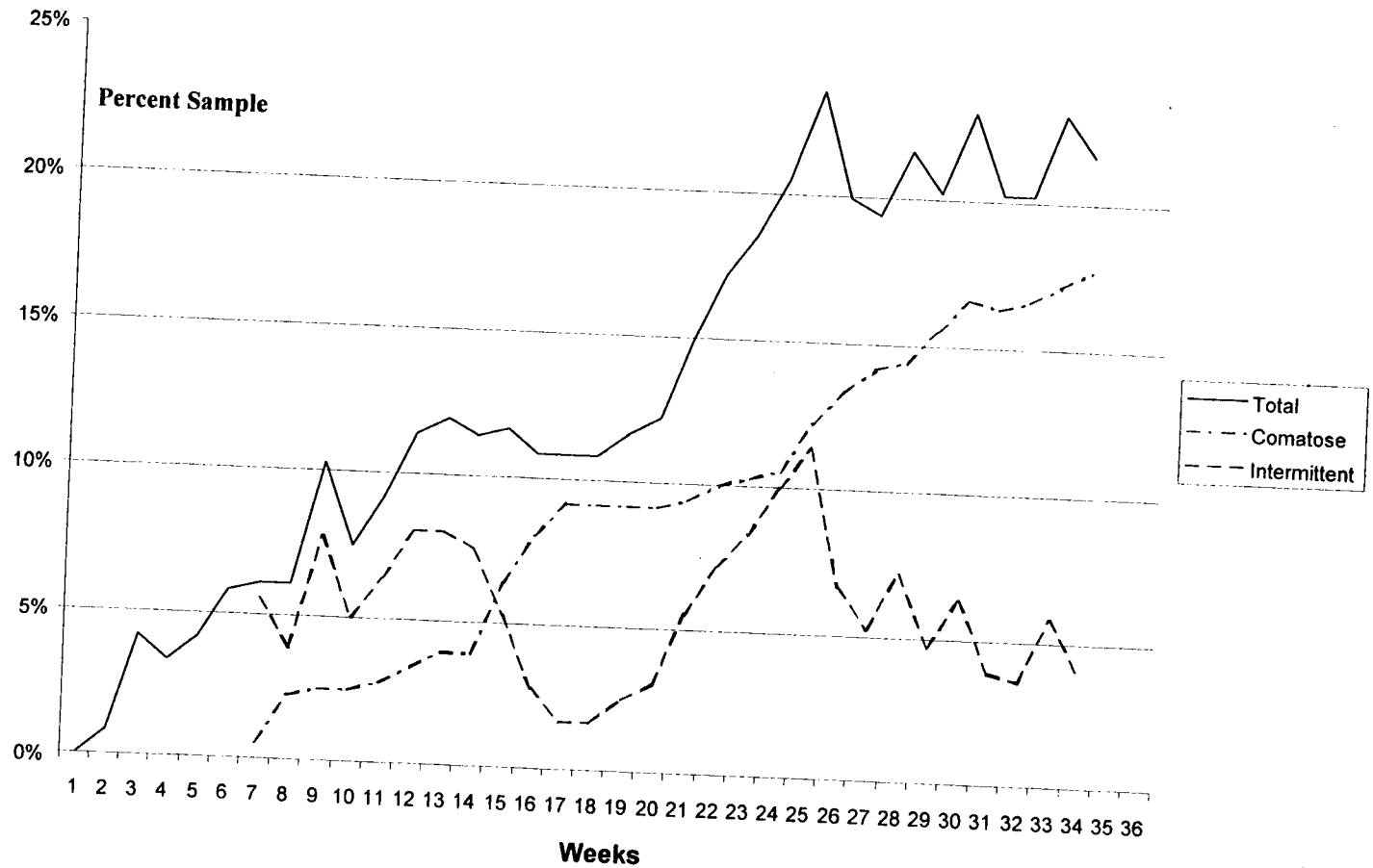

Figure 4.2 Web Page Persistence Over Time

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

Change of Address

Web pages as well as Web sites occasionally change URLs. Over the study period, 2.2 percent of the Web pages were 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 percent (two) became comatose shortly after the transfer, 37.5 percent (three) were modified significantly, and 37.5 percent (three) underwent no apparent structural or content changes. These Web pages were treated as comatose for purposes of subsequent analysis.

Original Web Site Size and Web Page Persistence

The size of the original Web site is not a particularly good indicator of Web page persistence (χ^2=17.2 p≤.245). There is, however, a tendency for Web pages on smaller Web sites to become comatose more often than those on larger Web are. Intermittence is also slightly less likely for Web pages on the larger sites and more likely for the smaller.

Inferred Domain and Web Page Persistence

Overall, inferred domain is not a particularly good indicator of Web page persistence (χ^2=18.7 p≤.175). Several domains are more or less stable than the rest. These are shown in Table 4.1.

TABLE 4.1 WEB PAGE INFERRED DOMAIN BY PERSISTENCE 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: χ^2=18.7 p≤.175

Government and organizational domains are less likely to become comatose than other domains. Government and military domains are 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.

If stability is considered the unlikelihood of coma, then government, organizational, educational, and military Web pages are more stable than average while commerical and unclassified geographic sites are less stable.

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 I have termed omega values (ϖ), 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 would be captured as "no change, " or zero.

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 ϖ_t, ϖ_c, ϖ_n, and ϖ_e. ϖ_t is the omega value for all forms of measured content and structural change. ϖ_c is the omega for content change. ϖ_n and ϖ_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 nine month Web page collection (first of January to end of August 1997), three percent of the sample returned an overall ϖ_t of zero. These eleven Web pages had no measured changes. At the same time, 0.8 percent of the sample changed in at least one respect each time, for an ϖ_t of one.

The sample average weekly ϖ_t varied greatly. Figure 4.3 is a plot of those values over the life of the project. Figure 4.3 contains two lines. The solid line plots the ϖ_t as a function of all time periods, whether the Web page responded or not. The dashed line is the same ϖ_t, except that it is calculated based on the number of periods any given Web pages 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, so parallel, subsequent analysis is based on a single omega function: periods present.

Figure 4.3 indicates a great deal of omega turbulence. At the onset of research, more than three-quarters of the sample exhibited some kind of change. This was marked by major omega swings during the first third. Variation persisted throughout the second and last third. By the last third, the swings began to smooth. This suggests that as Web pages mature, they become more stable over time.

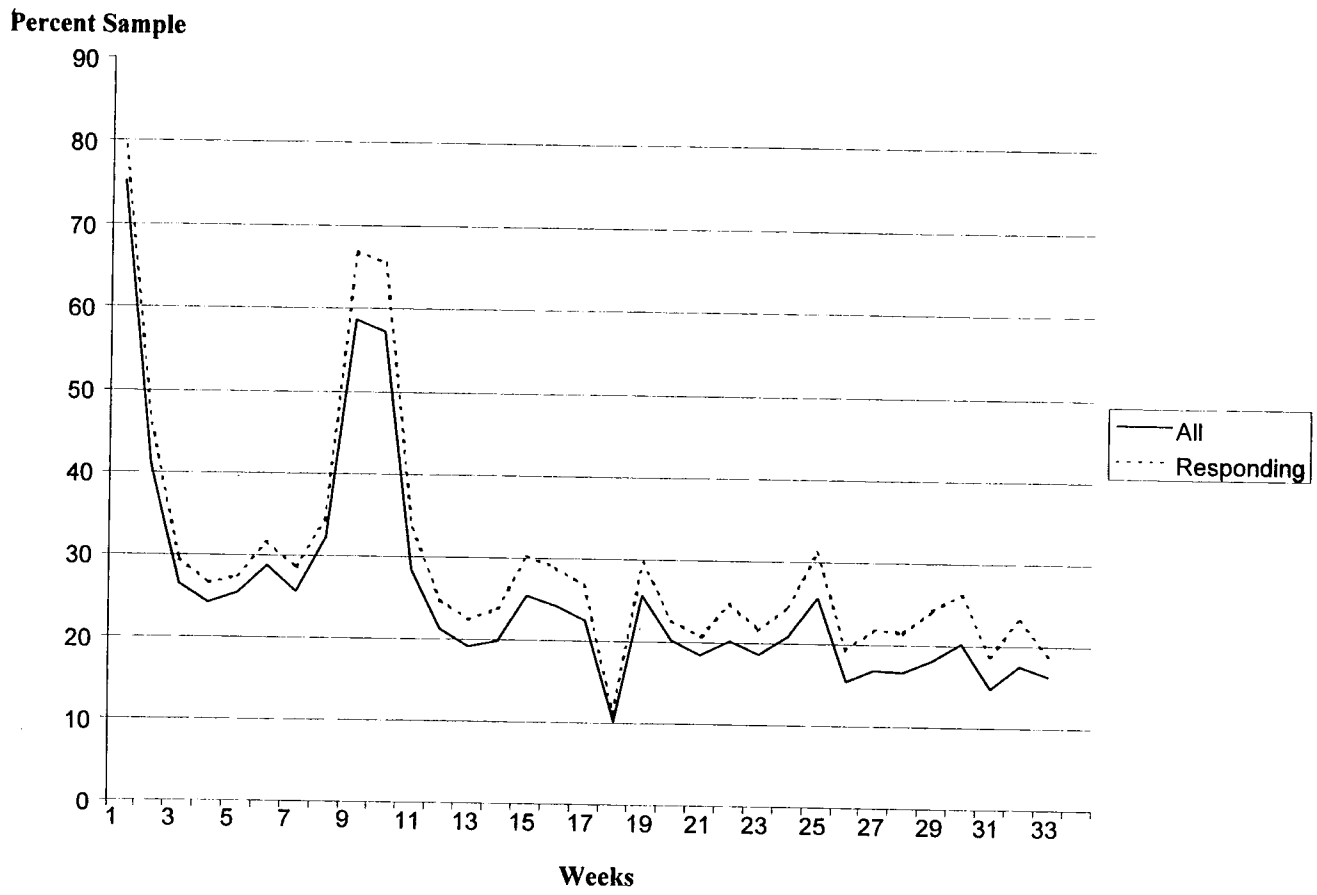

Figure 4.3 Web Page Content and Structural Change

A moderating and stabilizing factor could be that it is the more turbulent Web pages that drop away. The data presented in Table 4.2 do not support that conclusion. Table 4.2 offers the total sample ϖ_t, ϖ_c, ϖ_n, and ϖ_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.

TABLE 4.2 WEB PAGE OMEGA VALUES BY PERSISTENCE

Omega	Total Sample	Persistent	Intermittent	Comatose
ϖ_t	0.298	0.271	0.308	0.360
ϖ_c	0.239	0.234	0.221	0.287
ϖ_n	0.074	0.083	0.059	0.075
ϖ_e	0.150	0.176	0.127	0.120

Figure 4.4 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 ten and twenty percent of the sample during each period while new structural changes occur to between five and ten percent of the sample for the same period. Another possible explanation 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 twenty percent per week.

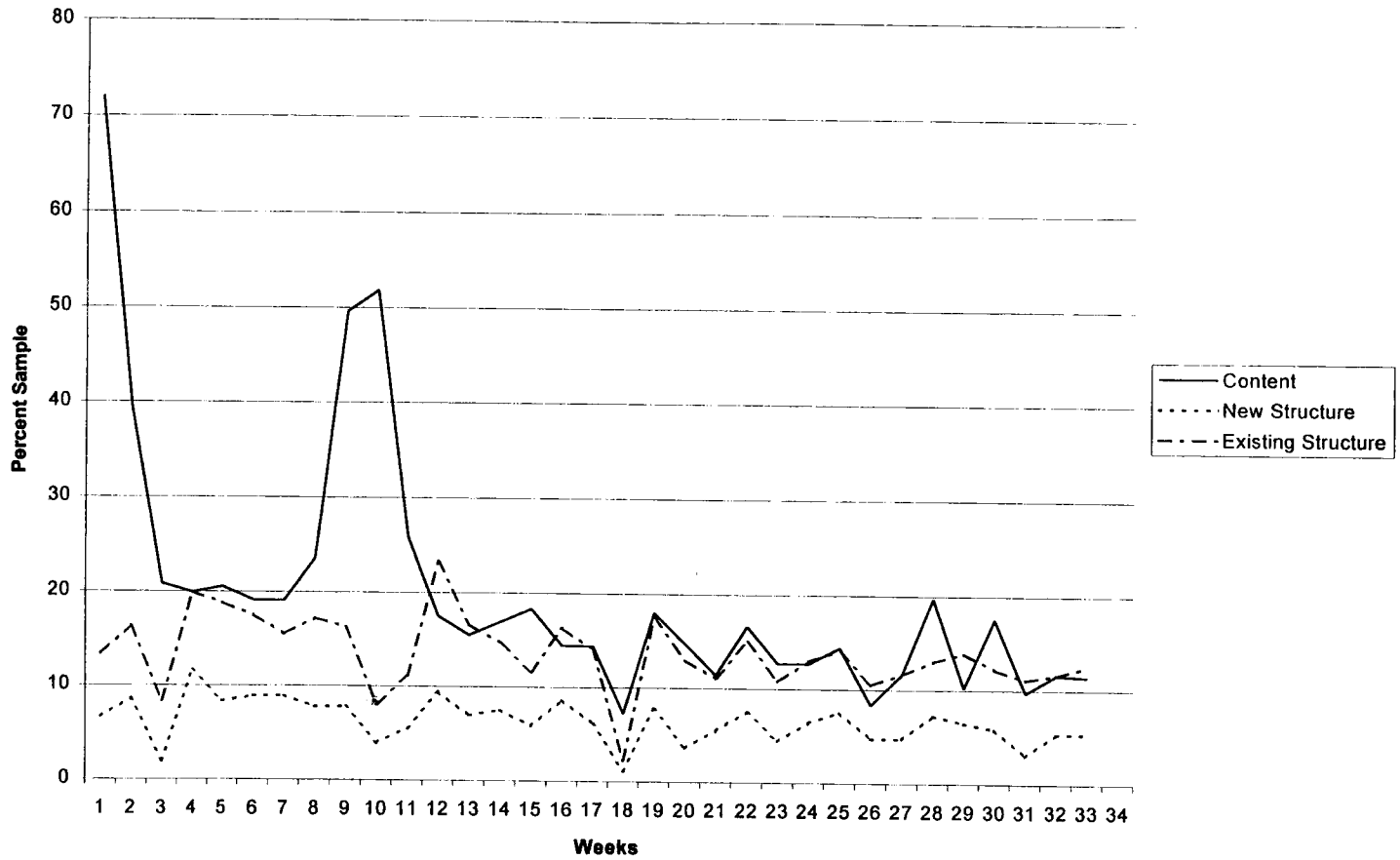

Figure 4.4 Web Page Change Components

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 4.3 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 contrarian conclusions. Both the smaller and big Web sites manifest Web page changes contrary to the expected.

TABLE 4.3 WEB PAGE OMEGA AND ORIGINAL WEB SITE SIZE

Omega	Total Sample	Smallest	Smaller	Small	Avg	Big	Bigger
ϖ_t	0.298	0.301	0.208	0.308	0.302	0.307	0.286
ϖ_c	0.239	0.253	0.182	0.259	0.226	0.247	0.239
ϖ_n	0.074	0.087	0.044	0.096	0.048	0.085	0.071
ϖ_e	0.150	0.166	0.098	0.156	0.142	0.165	0.128

Inferred Domain and Web Page Change

Web pages may vary according to their "publishers," to their different domains. Table 4.4 offers data according to the inferred domains (TLD, 2LD, and inferred). According to the data in Table 4.4, 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.

TABLE 4.4 WEB PAGE OMEGA AND INFERRED DOMAINS

Omega	Total Sample	com	edu	gov	mil	net	org	Uncl ISO 3166
ϖ_t	0.298	0.361	0.255	0.300	0.366	0.327	0.213	0.245
ϖ_c	0.239	0.295	0.186	0.214	0.384	0.300	0.105	0.148
ϖ_n	0.074	0.111	0.048	0.052	0.079	0.105	0.097	0.035
ϖ_e	0.150	0.226	0.105	0.167	0.182	0.170	0.105	0.081

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 six months are offered in Tables 4.5 and 4.6. Email 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. 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.

TABLE 4.5 WEB PAGE OMEGA AND FIRST WEB SITE OBJECT DOMINANCE TYPE

Omega	Total Sample	Average	Email	Graphic	Multimedia	Retriever	Text
ϖt	0.298	0.286	0.111	0.273	0.414	0.301	0.303
ϖc	0.239	0.228	0.122	0.222	0.357	0.234	0.247
ϖn	0.074	0.074	0.015	0.055	0.052	0.089	0.086
ϖe	0.150	0.140	0.051	0.130	0.269	0.152	0.149

TABLE 4.6 WEB PAGE OMEGA AND SECOND WEB SITE OBJECT DOMINANCE TYPE

Omega	Total Sample	Average	Email	Graphic	Multimedia	Retriever	Text
ϖt	0.298	0.294	0.186	0.358	0.291	0.286	0.259
ϖc	0.239	0.238	0.113	0.262	0.256	0.222	0.214
ϖn	0.074	0.080	0.036	0.092	0.069	0.065	0.061
ϖe	0.150	0.157	0.067	0.199	0.169	0.157	0.122

WEB PAGE PERSISTENCE AND CONSTANCY CONCLUSIONS

Web pages, like Web sites, demonstrate a significant variation in persistence and constancy behaviors. Like Web sites, the Web page sizes contract and constrict 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.

Three types of Web page longevity behaviors were identified: persistence, 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. That translates into a half-live of approximately one and a half years for Web pages. Web pages are also intermittent (as I suspect Web sites are as well). At any given time, approximately five percent of all Web pages are not answering the call, but will return.

Finally, 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 longevity 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.

Chapter 5

Conclusion

INTRODUCTION

This research had its beginnings in a need to better understand the dynamics of Web document behavior to meet a set of contract requirements. That contract called for the development of methodologies to aid in the development of a catalog of Web pages. We quickly realized that first, traditional intellectual content classification is not easily applied to Web documents. Second, we also grew to appreciate the heretofore-unappreciated complications for cataloging inherent in an ephemeral medium (McDonnell, Koehler, and Carroll 1998). This study is an outgrowth of those first efforts.

THE WEB IS DIFFERENT

World Wide Web materials are different from the "traditional," which include both print and other electronic media. There are at least seven significant differences between the two.

- First, as the findings presented in Chapters 3 and 4 bear out, WWW material is ephemeral and impermanent while traditional media are far more permanent.
- Second, WWW material is inconstant. It changes. Once published, traditional media do not change (or at least, not often).
- Third, authority and quality issues are, at least for the moment, vague on the WWW, more clear with traditional materials.
- Fourth, metadata are often uncertain or vague and often ignored on the Web. Systems are far more developed and accepted for traditional materials.
- Fifth, Web materials point to as well as incorporate other native Web materials both as bibliographic references and as integral parts. Traditional materials more often utilize bibliographic surrogates.
- Sixth, a collection of Web documents requires little or no physical maintenance. Libraries of traditional media require much.
- Seventh, catalog maintenance of Web documents requires much attention, traditional materials far less.

We should not bemoan the fact that the WWW is inherently different from traditional media. Nor should we throw our hands up in the face of what some might see as a paradigm shift and declare the present too complex to manage. We are not facing a paradigm shift. We are adding a new wrinkle, albeit a major wrinkle to information management, communication, and processing. Librarians and information scientists have not lost their mission. That mission has been expanded and, I believe, enhanced by the Internet.

NEW APPROACHES

We continue to work toward an understanding of the dynamics of the WWW as a medium for information transfer and use. I offer the following preliminary observations for consideration.

- The medium is no more the message any more than pen and paper is Hamlet or the word processor that this was written with is the study. Yet, first editions are collectors' items. There is beauty in presentation.
- We should not become too preoccupied with the technology, but rather with the role of that technology and the implications of that technology for communicating ideas. We need to be aware of our traditional and new roles as information scientists and librarians. We are concerned with providing access to the intellectual content of what we are charged to preserve. We are not changing what we do but how we do it.
- We need to conceive of Web sites and Web pages in ways other than as surrogates for the traditional. A Web page is not a page in the traditional sense of the word. Web pages consist of a series of Web objects. A Web page is like one of Alexander Calder's mobiles. Objects, "dingles" hang from a central object. Each object performs an individual task. At their base is usually but not necessarily a text object. A text object is a Web object that is created and edited using a specific set of tools based on a specific set of standards. The base object can be graphic object. That too is a Web object created and edited using a specific set of standards. A text object and a graphic object can carry the same message. Dingles are attached to the base object. Dingles can be added, eliminated, or changed at will. Manipulation of the dingles can and does impact the message as much as editing the base object. If the page is built well it has balance.
- The Web is different from traditional presentation media. Traditional implies that everything is "fixed" and immutable once published. A Web document is fluid. Once changed, the old ceases to exist. In the traditional world, the changes become new editions.
- All storage media are transitory. Some media are more transitory than others.
- Is it worth the trouble to try to develop bibliographic access to the WWW? Should we just fall back and rely on the search engines with their inadequate indexes to do a poor job of it?

We are challenged as never before to provide access to information in its several media. Many ideas have been offered, but none have yet been fully embraced. These ideas range from archiving the WWW (Kahle 1997) to providing permanent addresses for Web objects. FirstSearch's NetFirst classifies by content. We have suggested that Web pages can also be categorized by function (McDonnell, Koehler, and Carroll 1997). All are fraught with complexity.

A variety of solutions have been developed and will continue to be developed. These include sophisticated search engines and indexed directories. Other strategies are needed, strategies not solely reliant on keyword approaches. Several major search engines offer non-keyword search alternatives. These include the option to search on domains, on object content, by publication date or update, by language, by in-linkages, by URL, by directory structure level, and others. Web maps and link trekking can also serve as useful search tools (Koehler 1997a).

THIS MINOR CONTRIBUTION

The findings presented in Chapters 3 and 4 underscore the complexity of managing and categorizing the intellectual content of the WWW and providing efficient access to that content. This study has sought to offer two interrelated approaches to the study of the WWW.

First, recognizing the need for further Web classification work, I sought to develop Web site categories based on objective observable or measurable characteristics. The first set is URL based. Web sites and pages can be classified by their domain names. These can indicate functional and/or geographic publisher types.

URL elements can also be used. These include the transmission medium, the depth of the directory structure, the identification of alternative ports, and the use of the tilde to indicates levels of discontinuity.

Web sites, I argue, should be disaggregated according to points or levels of discontinuity for purposes of categorization and cataloging. It makes little sense to focus our attention at the server-level domain if that domain piggybacks an array of otherwise unrelated material.

The characteristics and the mix of the Web objects they contain can also categorize web sites. Web sites can be weighed. The number and type of objects can be counted.

While we all recognize that Web sites and Web pages change and change frequently, there have been few studies to date that address the phenomena of WWW change. With the exception of this one, I know of none that have approached Web change from the perspective of treating Web pages and sites as coherent wholes. I know of no studies other than this one that have attempted to understand changes based on Web site and page type.

Second, this study also sought to measure Web page and Web site change dynamics over time. The good news is that the rate of change fluctuations decline over time as Web pages and sites mature. The bad news is that the change is still very high for mature Web documents. Not only do Web pages and Web sites have relatively short half-lives, those Web pages and sites which persist undergo less but nonetheless frequent change.

A metric, the omega index, is offered as one measure of change. A standard metric can comparisons of Web documents. The measure can also provide a threshold for selection or non-selection of documents by librarians or catalogers.

SO WHAT . . .

The dynamics of Web change and the implications of that change on the creation, development, and communication of ideas are still unknown. This study was not designed to nor did it attempt to assess the impact of these changes on the intellectual content of Web documents. However, Web document extinction explicitly erases intellectual content from the Web and at least some change implicitly modifies

that content. This work has sought to clarify in a small way what the problems finding those answers can be.

This has been a limited work. It addresses Web changes over a period of less than one year. It explored changes for a relatively small population of Web sites and Web pages. While the population size may have been adequate to offer generalizations of the Web as whole, it was too small to allow statistically significant statements to be made of the subsections and categories that were found.

As always, further and expanded research is needed. That continued research is needed in at least three areas:

- We need to continue to monitor Web sites and Webpages. There are still many unanswered questions about constancy and permanence and the medium to long term implications of that change.
- This study did not address content. It merely sought to measure change in the mix of Web objects or "dingles" on Web sites and pages. Much change was found. There is a compelling need to tie dingle change to its impact on intellectual content. How much change matters? Can we find/use automated metrics to measure that change?
- If we can use automated metrics, how often do we have to apply them? This work begins to try to answer that question. The answer is "often," perhaps, with all due apologies to Douglas Adams, once every 42 days.

Bibliography

Bibliography

Brake, D. 1997. Lost in cyberspace. *New Scientist* 154 (2088): 12-3.

Bush, V. 1945. As we may think. *The Atlantic Monthly* 176 (1): 101-8.

Chankhuntod, A., P. Danzig, C. Neerdeals, M. Schwartz, and K. Worrell. 1995. A hierarchical Internet object cache. http://excalabur.usc.edu/cache.html. Updated November 6, 1995.

Claffy, K. 1996. "but some data is worse than others": Measurement of the global Internet. http://www.nlar.net/zachary.html. Dated June 1996.

Daniel, R. 1995. Uniform resource characteristics (URC). http://www.acl.lanl.gov/URC/. Dated November 3, 1995.

Daniel, R. n.d. Uniform resource identifiers (URI). http://www.acl.lanl.gov/URI/uri.html.

Daniel, R. 1996. Uniform resource names. http://www.acl.lanl.gov/URN/. Dated August 16, 1996.

Desai, B. C. 1997. Supporting discovery in virtual libraries. *Journal of the American Society for Information Science,* 48 (3): 190-204.

Dillon, M., and E. Jul. 1996. Cataloging Internet resources: The convergence of libraries and Internet resources. L. Pattie and B. Cox, eds., *Electronic Resources: Selection and Bibliographic Control,* New York: The Haworth Press: 197-238.

Duranti, L. 1989. Diplomatics: New uses for an old science. *Archivaria* 28 (1): 7-17.

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

Graham, P. 1997. Bibliography on electronic library issues. http://aultnis.rutgers.edu/texts/intpressbib.html. Dated January 27, 1997.

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

Kahle, B. 1997. Preserving the Internet. *Scientific American* 276 (3): 82-3.

Koehler, W. 1996. A descriptive analysis of Web document demographics: A first look at language, domain names, and taxonomy in Latin America," in Ching-chih Chen, ed., *Proceedings of the 9th International Conference, New Information Technology*, Pretoria, South Africa.

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

Koehler, W. 1997b. Document persistence on the WWW. *Proceedings 1997 Crimea Conference on Libraries and Associations in a Transient World.* Sudak, Ukraine.

Koehler, W. 1997c. Internet search note: Specialized retrieval and Web search engines. *Searcher* 5 (5): 63-5.

Koehler W. and L. Barnett. 1998. Domain name searching and World Wide Web search tactics. Forthcoming in *Searcher.*

Koehler, W. and D. Mincey, 1996. FirstSearch and NetFirst Web and Dial-up Access: Plus ça change, plus c'est la même chose. *Searcher* 4, (6): 24-8.

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

McDonnell, J., W. Koehler, and B. Carroll. 1997. Automating the dynamic development and maintenance of a distributed digital collection Forthcoming in. *Proceedings of the American Society for Information Science*, Washington, DC.

McDonnell, J., W. Koehler, and B. Carroll. 1998. Cataloging challenges in an area studies virtual library catalog (ASVLC): Results of a case study. Forthcoming in *Journal of Internet Cataloging* 1 (3).

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

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

Monk, T. and K. Claffy. 1996. A survey of Internet statistics/metric activities. *CCIRN, CIX Meetings*, Montreal, June 1996, http://nlanr.net/metricsurvey.html.

Monk, T. and k. claffy [*sic*]. 1996. Cooperation in Internet data acquisition and analysis. *Coordination and Administration of the Internet*, Cambridge, Massachusetts, September 8-10, 1996, http://www.tomco.net/~tmonk/cooperation.html#1.

Morgan, E. 1996. Possible solutions for incorporating digital data information mediums into traditional library cataloging services. L. Pattie and B. Cox, eds., *Electronic Resources: Selection and Bibliographic Control,* New York: The Haworth Press: 105-26.

NetCraft. 1997. http://www.netcraft.co.uk/Survey.

NetWizards. 1997. http://www.nw.com.

OCLC. 1997. PURL. http://www.purl.org, dated February 9, 1997.

Pattie, L. and B. Cox. 1996. Introduction. L. Pattie and B. Cox, eds., *Electronic Resources: Selection and Bibliographic Control,* New York: The Haworth Press: 1-8.

Pike, J. 1997. Shocked by search engine indexing. *Anchor Desk.* http://www5.zdnet.com/anchordesk/talkback_13066.html. Dated April 1, 1997.

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

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

Salamonsen, W., and R. Yeo. 1997. PICS-aware proxy system versus proxy server filters. *Proceedings of the Internet Society Meeting, The Internet: Global Frontiers*, CD-Rom, and http://www.isoc.org/isoc/whatis/conferences/inet/97/proceedings/A7/A7_3.HTM, Kuala Lumpur.

Tillman, H. 1997. Evaluating quality on the Net. http://www.tiac.net/users/hope/indqual.html. Revised April 17, 1997.

University of Waterloo, Scholarly Societies Project. 1997. URL-Stability Index for the Scholarly Societies Project. http://www.lib.waterloo.ca/society/URL_stability_index.html.

Urgo, M. 1997. 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.

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

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

Zakon, R. 1997. Hobbes' Internet Timeline v3.1. http://info.isoc.org/guest/zakon/Internet/History/HIT.html.

VITA

Wallace C. Koehler, Jr. received his primary and secondary education in Oak Ridge, Tennessee. He received his B.A. in political science from the University of Tennessee, Knoxville in 1971. He received his first M.A. again from the University of Tennessee in 1973. He then pursued doctoral studies in Government at Cornell University where he received a second M.A. in 1976 and the Ph.D. in 1977.

His daughter holds two degrees from the University of Tennessee. His son, at the time of writing, was a sophomore at the University of Tennessee.

After accepting the utter futility of recovering from two reductions-in-force in the political science arena, Dr. Koehler began the study of Information Science in 1975. This thesis culminates that effort.