Classifying Web sites and Web pages: the use of metrics and URL characteristics as markers

WALLACE C. KOEHLER, JR.

Wallace Koehler teaches information science at the University of Oklahoma. He received his PhD in political science from Cornell University in 1976 and the MS in information science from the University of Tennessee in 1997. His research interests include mapping WWW dynamics and the social science of information ethics.

Points to the way in which computer scientists and librarians working with the World Wide Web are turning to traditional library and information science techniques, such as cataloguing and classification, to bring order to the chaos of the Web. Explores cataloguing opportunities offered by the ephemeral nature of materials on the Web and examines several of the latter's unique characteristics. Suggests the coupling of automated filtering and measuring to the Web record cataloguing process, with particular reference to the ephemeral nature of Web documents and the ability to measure Uniform Resource Locator (URL) and Web document characteristics and migrate them to catalogue records using automated procedures. Reports results of an ongoing longitudinal study of 361 randomly selected Web pages and their Web sites, the data being collected weekly using the Flashsite 1.01 software package. Four basic approaches to ordering information on the Web were studied: postcoordinate keyword and full-text indexes; application of both precoordinate and postcoordinate filters or identifiers to the native document by either authors or indexers; use of thesauri and other classification schemes; and bibliometric techniques employing mapping of hypertext links and other citation systems. Concludes that off-the-shelf technology exists that allows the monitoring of Web sites and Web pages to 'measure' Web page and Web site characteristics, to process quantified changes, and to write those changes to bibliographic records. Capturing semantic or meaningful change is more complex, but these can be approximated using existing software.

INTRODUCTION

The World Wide Web (WWW) has brought new and fascinating challenges to the identification and management of information. The demands for management of the resource have brought together an ancient discipline – library science, – together with a new one – computer science, – in an interesting shotgun marriage of necessity that gives rise to their offspring – information science. In a somewhat irreverent commentary, one observer (Riddle, 1996) describes the shortcomings of librarians as:

> [Librarians] can never quite understand the technology as well as the people who invent it. While many librarians are struggling valiantly to keep up with runaway technology, they are continually in a position of reacting to rather than originating change.

And of computer scientists as:

> Computer people are continually trying to reinvent concepts which librarians have been honing for decades... [C]omputer people try to solve every problem by throwing robots at it.

The information access explosion has given rise to the need to manage that resource. Library scientists have been developing cataloguing schemes for almost as long as there have been libraries. Librarians are still somewhat bewil-

Address: School of Library and
Information Studies,
University of Oklahoma,
401 West Brooks, Room 120,
Norman, Oklahoma 73019-0528, USA
Tel: 001 405.325.3921
Fax: 001 405.325.7648
E-mail: w.koehler@ou.edu

dered as to how to bring their experience to the table, but that is changing. Computer scientists, on the other hand, have not been so reluctant to reinvent the classification and cataloguing wheel: 'Librarians who attend Internet developers' conferences ... sometimes refer to the discussions of cataloguing which take place there as 'library school kindergarten' (Riddle, 1996). We can summarize the differences between the two broad disciplines by suggesting that library scientists think analog to get to digital, while computer scientists think digital to get to analogue.

Some despair of ever organizing information and metadata for the WWW so that it can be effectively and efficiently used (e.g. Ardito, 1998). Many others recognize that the Web provides useful valuable information, but that access should be effectively organized using recognized and established library principles (Olsen, 1997). In an eloquent defense of Web cataloguing, Jul et al (1997) describe and dispel what are for them the three most persuasive arguments against the practice. These are:

- Web content is trash,
- Web content is too ephemeral, and
- cataloguing technologies were designed for print and are not applicable to the WWW.

Jul *et al* are correct. Yes, some of the Web is trash; some print is also trash. But even trash (a highly subjective subject) needs at times to be classified or catalogued. Their other two observations are more telling.

This paper addresses the latter two issues raised by Jul et al. The Web is ephemeral, but that can be used to advantage in cataloguing. Some 'traditional' cataloguing technologies should only be applied to print, but many others are appropriate for both. In addition, there are Web-based technologies that create cataloguing opportunities for the Web that are not appropriate for print. This paper explores cataloguing opportunities offered by the ephemeral nature of the Web and examines several characteristics unique to the Web.

It is an underlying assumption here that the marriage of the computer and library sciences can result in highly effective Web classification schemes and therefore information retrieval. It is not only possible but also desirable to couple automated filtering and measuring to the Web record cataloguing process. Automation is not the final product; it is a step to aid and augment, not to replace Web cataloguing and document retrieval. In the end, I believe, it is neither the librarian nor the computer scientist who will establish Web cataloguing schemes, it is rather the information scientist who will decide when 'throwing robots' is appropriate, when more traditional approaches are, and when the two should be mixed.

This paper briefly describes some of the efforts to manage the Web. I suggest that these efforts address some, but not all, aspects of Web documents that could, and therefore should, be used in the cataloguing process. The ephemeral nature of Web documents as well as URL and Web document characteristics can be measured and migrated to catalogue records using automated procedures.

APPROACHES TO WEB MANAGEMENT

The World Wide Web offers new and unique challenges to those seeking to bring bibliographic control to the medium. The information community has taken four basic approaches to bring some kind of order to the chaos of the Web. These are

- the development of post-coordinate keyword and full-text indexes;
- the application of both pre-coordinate and post-coordinate filters or identifiers to the native document by either authors or indexers;
- the use of thesauri and other classification schemes;
- bibliometric techniques employing mapping of hypertext links and other citation systems.

Examples of these are discussed below.

Commercial solutions include the many general robot-driven inverted index-based search engines like Lycos, AltaVista, Northern Light, Excite, HotBot, Open Text, WebCrawler, and so on. The early search engines offered keyword searches of title fields, but they have become far more sophisticated. Most online search engines now provide full-text searching, and a number provide non-text search capabilities. Several search engines index Web page keyword and abstract metatags. They also index date fields, author fields, as well as a range of domain, URL, object type, language, and other characteristics. Online and front-end meta-search engines have also been developed that submit searches simultaneously to multiple search engines. The more sophisticated provide filtering and automated search features.

It is the general consensus that these search engines in their present state provide valuable service but are not nearly so comprehensive or inclusive as their billing would have it (Tomaivolo and Packer 1996; Brake, 1997; Lawrence and Giles, 1998). There is also a growing number of limited area search engines (LASE) with specialized indexes based on subject or geographic criteria. These LASEs may include all Web material about or originating in a given area or about a given subject or they may maintain rigid selection criteria for collection inclusion.

The post-coordinate commercial directories like Yahoo! and OCLC's NetFirst are a different approach to providing bibliographic control to the Web. There has been a number of 'private' Web catalogues created based on these principles. They employ more traditional methods including abstracting, provision of multiple access points through the use of keyword and other descriptors, and standardized thesauri and classification systems. McDonnell, Koehler, and Carroll (1998), for example, describe one effort to develop a Web catalogue of area studies material, management of the ephemeral character of the collections, and their efforts to apply recognized standards to create records using the MARC template.

There are a number of systems employed to classify Web pages and Web sites. A number of information scientists are examining patterns among URLs, top-level domain names, site structures, page types, hypertext link types, and other indexable markers to identify content, quality, timeliness, and other factors (Chu, 1997; McDonnell, Koehler, and Carroll, 1998; Koehler, 1997a; Urgo, 1997). One approach classifies them according to function. McDonnell, Koehler, and Carroll (1997) describe jump pages, gateways, and content pages as well as variations. Gateway and jump pages are forms of navigational or architectural pages, analogous perhaps to catalogues and tables of contents, pointing to as well as containing information. Content pages contain the information the Web author has sought to disseminate.

There are numerous academic research projects and corporate research and development efforts to address the Web search and retrieval conundrum. Gerry McKiernan's Project Aristotle at Iowa State University (Project Aristotle, 1998) offers a continually updated list of projects together with brief descriptions and links.

Others have adopted bibliometric approaches by following hypertext link trails to 'authoritative' or substantive sites (Chu, 1997; Khan and Locatis, 1998). Recent work suggests that following link trails to clusters of clusters (Kaiser, 1998) can identify authoritative Web sites. Those documents identified as authoritative can be so tagged using a variety of existing methodologies. Purely quantitative evaluation methods are inadequate for they may result in the identification of popular rather than authoritative clusters. Quantitative cataloguing however sophisticated, still leaves room for quality control.

There are several indexing schemes adopted, in testing, under consideration, or proposed to provide additional bibliographic control to the Web. These include the Platform for Internet Content Selection (PIC) system, where Web authors may apply pre-coordinate alphanumeric codes as Web page descriptor metatags. Those metatags may also be applied later and interpreted by centralized or localized servers (Salamonsen and Yeo 1997). Dublin Core is another major proposal (Desai, 1997). The Dublin Core elements list and semantic headers represent a loosely constructed thesaurus in which index terms are author applied. The CyberStacks Project at Iowa State University is a demonstration project developed to apply Library of Congress classification to the Web (CyberStacks, 1998). NetFirst carries Dewey Decimal classifications.

Web libraries proliferate. Web libraries can be defined as Internet resident resources that point to, collect, categorize, and / or catalogue other Web resources. Web libraries range in size, scope, quality, and depth from rudimentary jump pages linking to pages with similar content to sophisticated, monitored, and specialized collections like the Social Science Information Gateway (SOSIG, 1998). Library selection and de-selection may be performed by the Web author alone, by teams of subject specialists and librarians, or by automated procedures. The smaller libraries typically organize material by subject, author, or some other criterion, and provide links to the collected document. The larger libraries often provide site maps as well as limited area search engines to facilitate access and retrieval functions.

Cognizant of Web site changes, at least one Web library has developed selection criteria that consider URL stability as one factor. The Scholarly Societies Project at the University of Waterloo (1998) has published to the Web a collection of URLs for professional organizations worldwide. They have found that 'canonical' URLs are more stable than the non-canonical. Stability is defined as longevity. Acanonical URL is one that contains the organization name without reference to other hosts or servers: e.g. www.orgname.org. They now have a preference for canonical URLs.

Both OCLC and the US Library of Congress have taken an active role in defining cataloguing criteria and rules for the Internet. These include defining and development of AACR2 metadata and documentation. The 1997 *Guidelines for the use of field 856* specifies the use of the 856 field and its subfields (Library of Congress, Network Development and MARC Standards Office, 1997a). The Library of Congress has also taken an active role in cross-mapping MARC, GILS, and the Dublin Core elements to facilitate the cross-fertilization and population of records across platforms (Library of Congress, Network Development and MARC Standards Office 1997b).

To implement these efforts and to develop Internet cataloguing, OCLC established Intercat, a catalogue of 'Internet-accessible materials.' Intercat is an experimental voluntary cataloguing effort supported by librarians. OCLC maintained the database and the access architecture (OCLC, n.d.). One fruit of this effort is OCLC's NetFirst, a part of the FirstSearch family of databases.

The Digital Libraries Initiative is a recent programme to develop and implement new electronic technologies to manage and utilize digitized data from a variety of sources (NSF / DARPA / NASA Digital Libraries Initiative Projects, 1998). The effort is to improve information access, retrieval, storage, processing, and transmission technologies. It is not to be a library of Internet materials but may include them. The research is funded by three United States Government agencies: the National Science Foundation, the Department of Defense Advanced Research Projects Agency, and the National Aeronautics and Space Administration. The research is to be performed at Carnegie Mellon University, the University of California at Berkeley, the University of Michigan, the University of Illinois, the University of California at Santa Barbara, and Stanford University.

EXPANDED CATALOGUING POSSIBILITIES

Laudable and useful as are all these efforts to manage the Web, they are inadequate on two grounds. The first 'inadequacy' is 'strategic.' Electronic and 'traditional' documents each have fundamental characteristics in common (Dillon and Jul, 1996). That said, we must recognize that the Web is a medium different from the traditional media. I have argued elsewhere that the Web does not represent an information paradigm shift. The Web is technologically sophisticated but as an information transmission and storage medium, it lies somewhere between the first type (centralized ownership but unrecorded) and the second type (recorded, distributed media). The Web contains elements of both; almost always, once the document owner erases the document, it ceases to exist, and once the owner edits that document, the intellectual provenance of the document is also lost (Koehler, 1998). The Web is truly ephemeral yet recorded and virtually universally accessible when available.

The second difference is tactical. There are Web page and Web site characteristics not shared by their more traditional type-two counterparts. These characteristics have been in large part not yet been adequately addressed either in the literature or in practice.

The data

The data reported in this paper result from an on-going longitudinal study of 361 randomly selected Web pages and the Web sites of which they are a part. Initial data were taken in December 1996. Web page data have been collected weekly using the software package Flashsite 1.01. The Web site data are collected using WebAnalyzer 2.0. Three Web site data harvests have been accomplished, each approximately six months apart. Both packages are products of Incontext. For further product information, see Incontext's Web site at *www.incontext.com*. Data management and statistical analysis were accomplished using Excel and SPSS. For an in-depth discussion of data collection and methodological issues see Koehler (1997b) and Koehler (1999).

Random selection of the Web pages was accomplished using WebCrawler's random URL generator. The sample was stratified to reflect the distribution of URLs by top-level domain in December 1996 (NetWizards, 1998).

The ephemeral Web

The electronic environment has and will continue to have an impact on libraries (Johnston 1998). That Web pages and Web sites are ephemeral is unquestioned. Web sites and Web pages undergo two forms of metamorphosis: persistence and change (Koehler, 1997b; Koehler, 1999?).

Document persistence

Persistence captures existence — is the page or site 'there' or is it not. Persistence is not concerned with changes to Web pages or Web sites other than presence or absence. A Web page or Web site may undergo significant 'alteration,' including the complete replacement of all Web objects on the document without affecting its persistence status. Those same Web pages and Web sites may also exhibit intermittence behaviour without change to the document once it returns.

Persistence may take one of three forms: permanence, 'comatoseness' and intermittence. A 'permanent' Web site or Web page is one that always resolves over some period. Acomatose Web page or Web site is defined as one that fails to respond or resolve after consecutive six weekly queries, including the most recent query. I prefer the term 'comatose' to 'dead' or 'defunct' because in theory and in practice a Web document once it 'disappears' can 'reappear' at the same URLif and whenever the Webmaster or Web author so chooses.

Web pages and Web sites appear to have half-lives of less than two and three years respectively. I have shown that the comatose rate for Web pages is virtually linear. By the end of the first sample year (December 1996 to January 1998) nearly 30% of the sample had disappeared. Web sites also disappear, although at a slightly reduced rate. By the end of the same sampling year, approximately 25% of that sample had 'gone' (Koehler, 1999). Not all that fail to respond or resolve at any given time are comatose. An intermittent Web page or Web site is one that has failed to respond at some point but that has returned. Web pages (and Web sites) can be classified according to their measured intermittence behaviour over a specified time period. These may be classified as *never, non-repeater,* and by *frequency*.

The majority of Web pages that persist do so without intermittence. However, a substantial minority are intermittent at least once and sometimes more often. At the end of the first data collection year 27.7% of the original 361 Web page sample was comatose. At any given time, between 2.5 and 10% of the sample is considered to be gone but intermittent; 4.2% were so deemed at the fifty-second collection. Of the non-comatose sample remaining after one year, 57.3% were always 'present,' thus 42.7% were at one time or another intermittently gone. Some are gone more often than others are. Of those intermittently gone over the year, 56.8% were gone for one duration only (one or more consecutive weeks), 30.6% were gone for two durations, 5.4% for three durations, 4.5% for four durations, and 2.7% for five or more durations. The intermittence duration ranged from one to 24 weeks. Most intermittences were from one to six weeks in length, and the average (mean) duration was 3.37 weeks for those present at the end of one year and intermittent at least once.

Document change

Document change offers a very different challenge. Over the sample year, almost all Web pages (more than 97%) and Web sites (more than 99%) underwent change. Change is defined here as variation in byte-weight, in object number, and/or the number and arrangement of hypertext links.

The purpose of this paper is not to offer interpretations of the meanings of that change, but to suggest that quantitatively measured change intervals and the relative degree of change can be used for bibliographic control. Often Web page and Web site change is trivial, but at other times it is substantial. The definition of trivial or substantial change is necessarily subjective. The addition of, say, a bandwidth demanding revolving logo may be taken as trivial by a reader seeking product price information, but as an important change by someone interested in Web site design. It should be noted that meaning can be changed by substituting one character for another, each with the same byte-weight, without impacting total byte-weight: consider for example the difference in meaning in these two phases 'the island was deforested' and 'the island was reforested.'

Web sites undergo significant change over time. Changes in the number of Web site total objects for the sampling period are shown Table 1.

Table 1. Web site relative total object changes in orders of magnitude in percentage

Order of Magnitude	Percent Dec/Jan 96/97 to June/July 97	Percent June/July 97 to Dec/Jan 97/98
Implode > -2 Orders of Magnitude	3.1	0.6
Implode > -1 < -2	3.7	6.2
Implode > 0 < -1	21.0	7.4
No Major Change	7.5	8.0
Explode > 0 < 1	50.2	68.5
Explode > 1 < 2	11.2	6.8
Explode > 2 Orders of Magnitude	3.4	2.5

Table 1 illustrates two general Web site change trends. First, Web sites are not static. They undergo dramatic size changes. Second, while Web sites both implode and explode, the general trend is toward size increases: well more than half of sampled Web sites increased in size over the sample year.

Web site catalogues are necessarily challenged by these size swings. As Web sites increase or decrease in size, sometimes as much as by more than two orders of magnitude, the depth and breadth of information contained in those Web sites are modified, amended, increased, or decreased. A dynamic index or catalogue must reflect those changes.

Web pages also undergo change. Two quantitative change measures can be identified. Changes in byte-weight (*content change*) and changes to the hypertext link structure of the page (*structural changes*). Byte-weight change is an imperfect surrogate for semantic or meaning change. Structural changes are modifications, additions, or deletions to the hypertext reference structure of the page. As such structural changes probably represent a subtler yet equally important source of semantic or meaning change as byte-weight.

These content and structural changes have been labeled as 'omega' (⌡). Omega can be divided into total omega ($⌡_t$) and three iso-types: content omega ($⌡_c$), new structural omega ($⌡_n$), and existing structural omega ($⌡_e$). Each omega value is an average of the periodic binary measure of change. The value '1' is assigned whenever dimension change of any magnitude occurs, '0' when none occurs. For example, if a Web page were to change one week but not the next, it would have an omega of 0.5 over the period. $⌡_t$ is not additive of the three iso-types. It reflects any change in any of the three measures. Thus, a Web page experiencing $⌡_c$, $⌡_n$ and $⌡_e$ changes would carry the same $⌡_t$ as one with only $⌡_c$ type change or other combinations.

Web pages undergo content change more frequently than structural. However, structural change may occur in the absence of content change. Mean omega values for the Web page sample at the end of the sample year were found to be $⌡_t$ 0.298, $⌡_c$ 0.239, $⌡_n$ 0.074, and $⌡_e$ 0.150 (Koehler 1999). Similar calculations could also be derived for Web sites.

Omega values can be employed in cataloguing in one of two ways. Individual Web page omega values can be calculated and included as part of the bibliographic record. These values can provide the user with an indicator of the rate and type of change the Web page undergoes. Omega values relative to the Web page population, the content of a digital Web library, or any other Web document pool can also be calculated. Thus, it may be useful to identify individual Web pages according to the degree to which they vary from some standard.

Web characteristics

There are a number of characteristics unique to the Web that can be used to develop bibliographic control. This paper explores two of those characteristics sets:

- bibliographic information which can be derived directly from the URL;
- and indicators based on quantitative evaluations of Web sites and Web pages.

URL-based characteristics

Web document uniform resource locators (URL) carry much useful information that can be used for Web page and Web site classification. URLs take the following general form: *transport medium://server.level.domain/directory/structure/....*

Transport media can be used for cataloguing purposes. The value of cataloguing by transport protocol is established. It is a subfield of the MARC 856 field ($2). The media include ‘gopher,’ ‘http,’ ‘ftp’ and other protocols. These not only designate transfer protocols, they also provide insight to document type and, in some cases, document quality. From one perspective, transport medium is not particularly useful to differentiate among Web documents since by far most URLs carry the hypertext transfer protocol (http) prefix. For example, in the sample more than 99% were http, and the protocol is growing even more dominant over time.

The second series of URLfragments provide address and domain information. The top-level domain (TLD) fragment is the right-most portion of the server-level domain (SLD). The SLD represents a form of address that resolves to the underlying Internet Protocol (IP) number. The IP number provides the actual Internet address for the server.

TLDs indicate ‘publisher type’ and take two general forms: functional and geographic. The functional TLDs include .com, .edu, .gov, .mil, .net, and .org plus seven more that may soon augment the .com TLD. Most functional TLDs originate (are ‘published’) in the United States, although there are a number of significant exceptions to the rule, particularly on the .com TLD.

The geographic TLDs are based on the two-letter ISO 3166 standard for country and regional abbreviations. They represent the country of publication. Examples include .au for Australia, .ca for Canada, .co for Colombia,.fr for France, .lk for Sri Lanka, .ru for Russia, .us for the United States, and .za for South Africa.

Many but not all geographic TLD registrars maintain second-level domain (2LD) functional tag practices (for a discussion of these and how they may be used in searching see Koehler and Barnett 1998). For example, .ac.uk represents academic servers in the United Kingdom, .co.jp indicates a commercial server in Japan, and .gob.mx is a Mexican government server.

Finally, it is also sometimes possible to ‘infer’ a functional domain from a geographic URL. Often universities, government agencies, and other entities will include the entity name in standardized and interpretable format. For example mcgill.ca, univ-lyon1.fr, unam.mx and leidenuniv.nl are academic institutions in Canada, France, Mexico, and the Netherlands while conicyt.cl is the 2LD for the National Commission of Science and Technology, a government agency in Chile. It is also sometimes possible to infer commercial domains from geographic URLs (e.g. *www.3m.cl* or *www.apple.de*) and to infer geographic from a functional URL (e.g. *www.republicofnamibia.com*, *www.nigeria.com*, or *www.guyana.net*

Web cataloguing schemes should include TLD and 2LD tags as well as inferred domains where appropriate. The distribution of TLDs in December 1996 included about 61% functional and 39% geographic domains (NetWizards 1998). This distribution is undergoing change from opposing directions. Many non-US based Web site owners prefer functional TLDs because these are seen to represent a more global presence. As a consequence, there is a trend toward migrating Web sites from geographic to functional root registrars. However, the number of new Web sites on geographic TLDs is increasing at a rate faster than those on the functional domain are.

Cataloguing by TLD, 2LD, and inferred domain can add additional granularity and therefore increased document differentiation. Table 2 illustrates increased sensitivity that can be achieved across the sample. Note that these percentages differ slightly from those reported for the WWW universe. They vary because of early attrition of Web sites in the original sample.

In Table 2, geographic domains are migrated to functional designations as 2LD and inferred domains are interpreted. Thus, the percent of geographic SLDs recognized as on educational or academic domains increases from 19.5% of the sample to 24.7 when 2LDs are included, and 29.9% when inferred domains are added. This strategy provides somewhat greater granularity because it increases functional identification of SLDs without decreasing the value of the functional / geographic TLD distinction.

Table 2. Impact of functional domain interpretation on geographic only granularity

Domain	TLD Only	TLD and 2LD	TLD, 2LD, and Inferred
Geographic	36.3	23.5	17.7
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
IP Number only	0.3	0.3	0.3

Web site structures

Web site organization can be mapped in one of two ways. The first is to map the relationship among Web page members of the Web site by hypertext links. The second is to map according to the server’s directory structure. These maps are very different and each provides useful data.

Hypertext maps include not only all Web pages located on the same server, they also include as part of the Web site those Web pages on other servers referenced from the Web site by hypertext link. These maps may also (but need not) include Web pages linked to the linked Web pages. *Reductio ad absurdum*, one could argue that almost all of the Web could be considered a part of each individual Web site through hypertext mapping.

Directory structure maps are limited to the arrangement of files as defined by the server operating system. Unlike hypertext maps, the directory structure can be read directly from the URL. Web pages at the top of the structure carry no structural indication and take the form *http://aaa.bbb.ccc*. These can be said to be at the 'zero level.' Web pages at the 'one level' take the form *http://aaa.bbb.ccc/xxx*. Each subsequent forward slash adds a level to depth.

Directory structure awareness can be important in the initial development of Web page catalogues and in defining the level or complexity of the cataloguing effort. I have found that Web pages located at the zero or one level are both about twice as likely to persist but also to change than are those Web pages located deeper on the directory structure (Koehler, 1999). It has been shown that Web pages closer to the zero level are more likely to be navigational, while those slashed lower are more likely to carry 'content' (McDonnell, Koehler, and Carroll, 1999). Zero level catalogues of navigation pages are likely to be persistent but contain changing records. Those catalogues of 'lower' content pages will not map changing records, but the underlying native documents will be less persistent.

An analysis of directory structure file names may prove useful in search and retrieval efforts if not in cataloguing. Urgo (1996) has found semiotic patterns in the file naming schemes on the .com domain. She argues that Web site designers often choose file names that reflect the file content. Examples include 'products,' 'ceo,' 'financial,' and so on. These file names, she suggests, might be captured as index terms.

The field of diplomatics may also contribute to the interpretation of both hypertext and directory structure maps. Diplomatics is concerned with the meaning that can be derived from an analysis of the form or appearance of a document (Duranti, 1989). Both types of Web site maps vary significantly in form and appearance. Both reflect differences in how information is organized and presented by Web authors. Different organization types may reflect differences in content, quality, authority, timeliness, and other variables of importance to information cataloguers and consumers.

Quantitative characteristics

Cataloguers frequently document various physical characteristics of the material for which they provide bibliographic control. These characteristics may include number and type of pages, frequency and type of illustrations, dimensions, binding, as well as other features. Some of these characteristics are always captured; others are only documented if their appearance is unusual or non-standard.

Web pages and Web sites can be characterized by their 'physical' attributes. Some of these characteristics are analogous to print, others are unique to hypertext documents. There is software available usually marketed for Web site diagnostics that can be used to provide quantitative measures for bibliographic control. The software used to capture data for this paper is WebAnalyzer 2.0, a product of InContext (*www.incontext.com*). WebAnalyzer and similar software packages can be used to measure Web page and Web site object mix, size, and hypertext link depth and density.

Object mix

Web sites consist of a collection of Web pages with related meaning or themes specified by the Web site author located on server level domains but also incorporating Web pages often authored by others on other SLDs.

Most SLDs consist of a single Web site (more than 80% in the sample). However, a single SLD may host several Web sites. These are often distinguished by discontinuity markers. Discontinuity markers include directory structure naming and the use of tildes.

Web pages consist of a collection of Web objects. Most Web pages have a text document as their base and any number of hypertext attached Web object arrayed from the base object much in the same way as one of Alexander Calder's mobiles. The number and type of Web objects are growing. These objects include text, graphics, audios, videos, mail, Java, ftp, gophers, and so on. These can be reduced to five main types: text, graphic, multimedia, file retrieval, and mail.

Web object types are not evenly distributed within Web sites. First, some Web objects are more frequently found closer or further from the analysed propositus page. The term propositus is borrowed from the genealogical lexicon and refers to the individual upon whom a genealogy is build both forward and back. For example, mail, file retrieval, and multimedia objects are often further away on a hypertext basis from the propositus than are text and graphic objects. This finding may contribute to a diplomatics analysis.

Second, the average (mean) number of each type of Web object varies, as is shown in Table 3. On average, Web sites consist primarily of text and graphics objects. The typical Web site in late 1996 contained less than 1% multimedia, file retrieval, and mail objects combined.

Table 3 provides a standard by which individual Web sites can be classified. Variation from each of the 'average' number of Web objects can be measured and catalogued. Each of the five categories can be divided into ordinal classes: for example from low to high text, graphic, multimedia,

Table 3. Web site percentage Web object means and standard deviations, December/January 1996–97

Web Object Type	Mean	Std. Dev.
Text	56.1	21.8
Graphic	33.8	215
Multimedia	0.078	0.42
File Retrieval	0.17	0.53
Mail	0.76	12.6

file retriever, or email content. Each class can be based on individual Web object means and standard deviations. I have suggested one approach (Koehler, 1997b) that results in classifying Web sites according to their dominant Web objects and variation from the 'average' model. Based on a set of ordinal categories for the percent of Web objects, Web sites can be reduced to six general types:

- Average (no dominant Web object)
- Wordsworth (text dominant)
- Coffee-Table (graphics dominant)
- Mogul (multimedia dominant)
- Retriever (ftp / gopher dominant)
- Post Office (e-mail dominant).

The names follow WWW naming practices; they are relevant but also slightly irreverent. In December 1996 and January 1997, the object dominant sample was Average 41.0%, Wordsworth 21.2%, Retriever 17.4%, Coffee-Table 13.1%, Mogul 6.1%, and Post Office 1.2%.

The object dominance standard reported here or one like it can be generated using commercially available off-the-shelf software. Individual Web site Web object distributions can be automatically calculated and those statistics can be migrated to Web document catalogues. It must be noted that as Web sites and Web pages change, so do their Web object mixes. Individual Web sites should therefore be reassessed periodically.

Web site and Web page size

Web site and Web page size can be used to classify Web documents. Web site size can be assessed in one of three ways: byte-weight, number of objects, and the number of Web pages in a Web site. The number of Web pages on a Web site is analogous to the number of text objects on the site. In addition, the number of hypertext links from various Web pages within a Web site to other pages on the site (internal links) and the number of links from a Web site to other Web sites (external links) can be measured.

The average number of objects and the byte-weight for each object class per Web site are shown in Table 4. As the statistics indicate, the size and construct of Web sites vary greatly. The number of text objects ranged for the sample reported in Table 4 from one to more than 13,000. The total number of all objects ranged from two to over 14,000. The total byte-weight of Web sites ranged from 292 bytes to over 52 megabytes (exclusive of audios and videos, which averaged 12 megabytes each).

Web site size can be classified in any number of ways. Because the byte-weight of individual Web objects vary both among types as well as from one like object to another, it may be useful to combine both byte-weight and object count. To achieve that end, byte-weight and object count values can be normalized, summed, then ordinal values assigned to the range. In this case (Koehler, 1997b), an average value (4) was assigned to all Web sites with values plus or minus 0.5 standard deviations from the mean. For each increment of one standard deviation from average, the assigned value was increased or decreased by one. This resulted in a size range from 1 to 6 or 'smallest' to 'bigger.' None qualified as 'biggest.' The sample ranged in size from 'smallest' 3%, 'smaller' 7.6%, 'small' 24.7%, 'average' 23.8%, 'big' 34.7%, and 'bigger' 6.2%.

Table 4. Web site size averages December 1996 and January 1997

Web Object	Mean	Median	Standard Deviation
Number of Objects			
Text	564.3	106	1472.4
Graphics	181.4	52	314.5
Audio	4.8	1663	37.33
Video	0.5	0	3.5
FTP	12.6	0	89.5
Gopher	6.3	0	21.9
Mail	61.4	4	240.2
Total	833.1	217	1712.3
Byte-Weight			
Text	1,360,733	222,829	3,336,292
Graphics	2,174,769	465,473	4,733,840
Total	3,539,064	977,203	6,480,651

Web page size is most usefully measured in byte-weight. Like Web sites, Web pages vary greatly in size. Individual Web pages range in size from zero kilobytes (kb) to more than 2000 kb, with no theoretical upward limit. Web pages also tend to increase in size over time. Data have been collected weekly since January 1997 and over the period ending on May 22, 1998, the average byte-weight of a non-comatose Web page increased from 58.66 kb to 111.17 kb, resulting in an annualized increased 'byte-creep' of more than 34%. Much of this is growth by accretion, new material is added to the Web page, while the older content is edited but not removed.

Over the same period, the number of non-responding Web pages (both comatose and intermittent) increased from zero to 43.8% of the sample. Thus, in an aging Web page collection, the size of the collection decreases over time measured by the number of non-comatose Web pages. But for those Web pages extant at any given time, their individual, average byte-weight tends to increase. It may therefore be desirable to indicate not only the intermittence rate of extant Web pages, but their individual growth rates as well. This is particularly true if byte-weight change is an imperfect yet useful surrogate for content change.

Link depth and density

Web site structures can be described by their directory structures as is discussed above. They may also be analysed according to the hypertext links to and from Web site pages as well as according to the density of Web pages by distance from the propositus page. The hypertext structure is presented as concentric rings. Those Web pages on the first ring have immediate hypertext ties from the propositus to themselves. Those on subsequent concentric rings are connected from the propositus through intermediate pages and are not directly linked from the propositus. It is both possible and likely that Web pages are connected through more than one route. Ring depth is determined as the most direct route. Density is a measure of the average number of Web objects or byte-weight on each ring on the Web site or from the propositus. Web site densities are measured here from the Web site homepage or index page. Home- or index pages are those which resolve from the SLD only or from an identified point of discontinuity. Ring depth and density include all Web pages at a Web site that are located on the same SLD. Web pages are included as part of a Web site that are not located on the SLD when and only when they are linked directly to one of the pages on the SLD. Thus, for this analysis, links from a non-SLD Web page are not included within the hypertext structure of a Web site unless they are directly linked from another qualifying page. Such a page might be linked to the propositus on the third level because it is directly linked to an on-SLD page linked to the propositus at the second level.

The number of hypertext Web site levels varies. The minimum number of levels, including the propositus, encountered in the sample is one, the maximum fifty-nine. There is no limit to the number of levels possible; the most the author has measured is 179. This 'lord of the rings' is an English university Web site. Most Web sites do not exceed eight levels. The December 1996 to January 1997 sample mean is 4.82, the median 4, and the standard deviation 4.88. Web site levels increase over time. In July and August 1997, the mean had increased to 5.22 levels, the median to 5 and the standard deviation to 4.67. Densities are 626.3 megabytes and 159.4 Web objects per level in the first period and 854.8 megabytes and 231.2 objects per level in the second.

The number of ring levels and densities provide structural data. Structure infers the organization of information. That, in turn, may provide insights into the importance of one 'piece of information' over another, the emphasis placed by the Web author on priority and order, as well as suggest information groupings or clusters. Further research is necessary to establish whether different ring and density configurations indicate different information qualities, quantities, authority, or other pertinent considerations. At minimum, individual Web site ring counts and densities can be measured and reported. Further work can establish whether there are different configuration types of significance to the library community.

CONCLUSIONS

Library scientists have advanced the field of cataloguing well beyond the stacking of scrolls by size or the use of playing cards to signal important documents. The complex and sophisticated card catalogue has evolved into an even more complex electronic management tool. It has evolved in part because the number of items and therefore records continue to increase at nearly geometric rates. The electronic tool is also more attractive because it offers multiple and manageable access points, which, because of physical limitations, a card catalogue cannot.

The catalogue is also changing because the character of the document is also changing. Digital documents can be stored, accessed, downloaded, and edited from afar. The World Wide Web further complicates the role and function of the catalogue. The doubly dynamic nature of native Web documents creates an inherent need for dynamic bibliographic control of those documents.

Three approaches to the management of Web documents have been suggested here. The first two have 'traditional' analogs. The first approach is to use elements present or inferred from Web document URLs for cataloguing. These include the transfer media, terms used to label files, as well as domain names from the top-level through the full server-level domain name. In 'traditional' cataloguing, title, publisher, author, copyright, and myriad other information might be said to be taken from any work's front material.

The second approach is to utilize data derived from the quantified measurement of Web sites and Web pages. These data include size, object mixes, and hypertext depth and density. Traditional cataloguing captures page and illustrations counts, document structure, and similar information.

The third is perhaps the most radical because, unlike the other two approaches, there are no parallels with the 'traditional.' It is concerned with the persistence and change of Web documents. Persistence and change can be measured. That data can be used to identify and classify Web documents. It is certainly true that cataloguers document changed or new editions, but almost always the new editions stand along side the old. This does not now occur on the Web, although efforts to archive it may mitigate this statement somewhat sometime.

The World Wide Web represents a challenge to those seeking to manage its information content. Estimates of the size of the Web vary greatly. At this writing, there were probably some 1.5 million server-level domains with some 2 million Web sites. Data for the number of Web pages are even more imprecise and estimates range between 100 and 600 million. To further complicate the situation, these numbers continue to grow.

The size and complexity of the World Wide Web and its document mix lend themselves to automation. Off-the-shelf technology exists that allows us to monitor Web sites and Web pages, to 'measure' Web page and Web site characteristics, to process quantified changes, and to write those changes to bibliographic records. Capturing semantic or meaning change is more complex, but again these can be approximated using existing software.

These processes are, in the end, no substitute for human judgement. They do, however, provide access points, filters, and flags for the cataloguer, the information scientist, and the end user.

June 1998

REFERENCES

Ardito, S. (1998) The Internet: beginning or end of organized information? *Searcher,* **6,** (1), 52-7

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

Chu, H. (1997) Hyperlinks: how well do they represent the intellectual content of digital collections?' *Digital Collections: Implications for Users, Funders, Developers and Maintainers, Proceedings of the American Society for Information Science,* **34**, 361-72

CyberStacks (1998) *http://www.public.iastate.edu/~CYBER - STACKS/homepage.html*

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

Dillon, M. and Jul, E. (1996) Cataloguing Internet resources: the convergence of libraries and Internet resources. In: L. Pattie and B. Cox, (eds) *Electronic resources: selection and bib - liographic control*. New York: The Haworth Press, 197-238

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

Johnston, C. (1998) Electronic technology and its impact on libraries. *Journal of Librarianship and Information Science,* 30, (1), 7-24

Jul, E., Childress, E. and Miller, E. (1997) '42: Don't panic, it's a common disaster' and '42: now that we know the answer, what are the questions?' *Journal of Internet Cataloguing,* **1**, (3), *http://jic.libraries.psu.edu/jic1nr3-42.html*

Kaiser, J. ed., (1998) New search strategy untangles the Web. *Science,* **280**, (5364), 647

Khan, K. and Locatis, C. (1998) Searching through Cyberspace: the effects of link display and link density on information retrieval from hypertext on the World Wide Web. *Journal of the American Society for Information Science,* **49**, (2), 176-82

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/inet97/proceedings/D3/D3_3.HTM*

Koehler, W. (1997b) Web site and Web page persistence and change: a longitudinal study. MS Thesis, The University of Tennessee, Knoxville

Koehler, W. (1998) The librarianship of the Web: options and opportunities managing transitory materials, *Proceedings 10th New Information Technology Conference*, Hanoi, Viet Nam, March 1998. (Ching-chih Chen, ed.), West Newton, MA: MicroUse Information

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

Koehler, W. (1999) An analysis of Web page and Web site constancy and permanence. Forthcoming in *Journal of the American Society for Information Science*

Lawrence, S. and Giles, C. (1998) Searching the World Wide Web. *Science,* **280**, (5360)

Library of Congress, Network Development and MARC Standards Office (1997a) *Guidelines for the use of Field 856*. Revised August 1997. *http://lcWeb.loc.gov/marc/856guide.html*

Library of Congress, Network Development and MARC Standards Office (1997b) *Dublin Core/MARC/GILS Crosswalk. http://lcWeb.loc.gov/marc/dccross.html.*

McDonnell, J., Koehler, W. and Carroll, B. (1997) Automating the dynamic development and maintenance of a distributed digital collection. Digital collections: implications for users, funders, developers and maintainers. *Proceedings of the American Society for Information Science.* Medford, NJ: Information Today, 244-59.

McDonnell, J., Koehler, W. and Carroll, B. (1999) Cataloguing challenges in an area studies virtual library catalog (ASVLC). *Journal of Internet Cataloguing,* **1**, (4) [forthcoming]

NSF/DARPA/NASA Digital libraries initiative projects (1998). *http://www.cise.nsf.gov/iis/dli_home.html*

OCLC (n.d.) Building a catalog of Internet-accessible materials. *http://www.oclc.org/oclc/man/catproj/overview.htm*

Olsen, N. (1997) Cataloguing Internet resources: a manual and practical guide, 2nd ed. *http://www.oclc.org/oclc/man/cat - proj/overview.htm*

Project Aristotle (sm): Automated Categorization of Web Resources (1998). *http://www.public.iastate.edu/~CYBER - STACKS/Aristotle.htm*

Riddle, P. (1996) Library culture, computer culture, and the Internet haystack. *http://is.rice.edu/~riddle/dl94.html.*

Salamonsen, W. and Yeo, R. (1997) PICS-aware proxy system versus proxy server filters. *The Internet: Global Frontiers, Proceedings, Internet Society Annual Meeting.* Kuala Lumpur, CD-ROM

SOSIG (1998) Social Science Information Gateway. *http://sosig.esrc.bris.ac.uk/welcome.html*

Tomaivolo, N. and Packer, J. (1996) An analysis of Internet search engines: assessment of over 200 search queries. *Computers in Libraries*, **16**, (6), 58-62

University of Waterloo, Scholarly Societies Project (1998), URL-stability index for the Scholarly Societies Project. *http://lib.waterloo.ca/society/URL_stability_index.html.*

Urgo, M. (1997) Analyzing company Web sites. *InfoManage*, **4**, (3), 7