Building Sustainable Collections of Free Third-Party Web Resources

by Louis A. Pitschmann

June 2001

Digital Library Federation
Council on Library and Information Resources
Washington, D.C.

About the Author

Louis A. Pitschmann is the incoming dean of libraries at the University of Alabama. At the time of this report's publication, he was completing a fifteen-year tenure as associate director for collection development and management at the University of Wisconsin-Madison. Before that, he worked for almost a decade at the Cornell University Libraries. Mr. Pitschmann received his Ph.D. and M.L.S. from the University of Chicago. He has presented papers and written several articles on various aspects of collection development and management in research libraries.

First Digital Library Federation electronic edition, September 2008

Originally published in trade paperback in the United States by the
Digital Library Federation and the Council on Library and Information Resources, Washington, D.C., 2001

Digital Library Federation ISBN-13: 978-1-933645-08-7

www.diglib.org

Contents

Preface

In January 2000, the Digital Library Federation (DLF) launched an informal survey to identify the major challenges confronting research libraries that use information technologies to fulfill their curatorial, scholarly, and cultural missions. With astonishing unanimity of opinion and clarity of voice, respondents pointed to digital collection development as their single greatest challenge. Whether the digital information came from a commercial publisher or from a digitization unit within the library, it seemed to exist under a cloud of profound and unsettling uncertainty. Would it be useful and useable in its present or intended form, or require additional work on the part of catalogers, systems staff, or subject bibliographers? What new demands would its availability make on library reference staff? What level of continued investment would be necessary to ensure its accessibility on current hardware and software?

The survey also revealed that leading research libraries had learned a great deal about their digital collections through experience. Though substantial, that learning had rarely been expressed outside the collection policies, working papers, and implementation guidelines that libraries create to coordinate and manage their collection development efforts. Accordingly, in April 2000, the DLF commissioned three reports to address broader concerns about digital collections. The three reports deal respectively with commercial electronic content, digital materials created from library holdings, and Web-based "gateways" that link to selected Internet resources in the public domain. The reports mark a starting point for what we hope will emerge as an evolving publication series.

Working to a common outline and based on learned experience, the authors demonstrate how decisions taken by a library when acquiring (or creating) electronic information influence how, at what cost, and by whom the information will be used, maintained, and supported. By assembling and reviewing current practice, the reports aim where possible to document effective practices. In most cases, they are able at least to articulate the strategic questions that libraries will want to address when planning their digital collections.

In this report, Louis Pitschmann deals with the widespread practice of listing "useful" Internet resources. Variously billed as subject gateways, Internet resource guides, and "related Internet resources," these inventories appear on library Web pages across the country. Their construction has become a cottage industry, often fuelled by the voluntary and devoted effort of individuals who have taken it upon themselves to identify and catalog worthy resources that occupy some definable segment of the World Wide Web. The redundancy involved in this work is as substantial as its long-term hidden costs. The author's treatment of the practice is critical, yet fair. He has few doubts about the value that such Internet resource guides offer to library users. At the same time, he asks searching questions about whether they may be developed and maintained outside the mainstream of collection development efforts and without the resources that typically support such efforts. Drawing extensively on evolving international experience, Mr. Pitschmann's work is essential for any institution that seeks to build its collection in part through reference to "free" Internet resources.

Daniel Greenstein
Director, Digital Library Federation

Acknowledgments

The author wishes to acknowledge the advice offered by colleagues, especially Susan Barribeau, Jo Ann Carr, and Barbara Walden of the University of Wisconsin-Madison, and Michael Seadle and Amy Tracy Wells of Michigan State University, who provided important leads and suggestions in the early stages of the project. Bonnie McEwan of The Pennsylvania State University provided much welcome advice on the penultimate draft of this paper. Tim Jewell, head of collection management services at the University of Washington, and Abby Smith, director of programs at the Council on Library and Information Resources, are owed a special debt of thanks for their unfailing moral support and encouragement throughout the project. The author also wishes to thank Daniel Greenstein, director of the Digital Library Federation, without whose guidance and counsel this report would not have been possible. Most importantly, the author wishes to acknowledge his wife's patience and understanding while this report was researched and written.

1. INTRODUCTION

The purpose of this report is to identify and synthesize existing practices used in developing collections of free third-party Internet resources that support higher education and research. A review of these practices and the projects they support confirms that developing collections of free Web resources is a process that requires its own set of practices, policies, and organizational models. Where possible, the report recommends those practices, policies, and models that have proved to be particularly effective in terms of sustainability, scalability, cost-effectiveness, and applicability to their stated purpose.

Background work for this report confirmed John Kirriemuir's observation that the selection of free third-party Web-based resources in North American academic libraries remains largely the responsibility of individuals working alone (Kirriemuir 1999). This background work also revealed that tasks ranging from identification of resources to their delivery through a library OPAC, Web page, or portal are seldom documented in a manner that would permit others to build upon existing practices, regardless of whether those practices demonstrate poor choices or good practices. This fact obtains whether one is looking for criteria used by an individual librarian or adapted by a library as institutional policy. For example, when interviewed about his Web selection criteria, one area studies librarian responded that he selected only "quality sites." When asked to define "quality," he replied, "We all know how to select good books, and Web resources are no different." Third-party public domain Web resources, however, are fundamentally different from scholarly print and analog formats as well as from commercially produced digital resources.

This report outlines the similarities and differences between print and free Web resources and describes how the nature and complexity of free Web resources comply with or challenge traditional library practices and services pertaining to analog collections. The report also recommends, where appropriate, certain practices that have proved effective in meeting desired goals in specific contexts. There is no single set of preferred or "best" practices that one should

follow in collecting free Web resources. What succeeds in one context might prove fruitless in another. The practices recommended in this report have proved effective in a specific project or show promise for broad applicability, regardless of the fiscal, technical, or human resource parameters in any specific institutional setting. Ultimately, it is not the practice per se, but rather how well that practice lends itself to a particular set of goals, workflows, and staffing preferences, that determines its effectiveness and value in any given setting.

1.1. Methodology

Because of the paucity of formal documentation, writing this paper required the use of several methods of data gathering.

1. *Interviews*. Librarians were asked to identify sites they considered to be "high quality" and to describe their criteria for evaluating the quality of a site. Every attempt was made to avoid implying or providing a definition of "high quality" when eliciting the librarians' opinions. This methodology proved useful for several reasons. First, it demonstrated the degree of subjectivity in defining this term. It made it possible to identify a large number of sites in the humanities and social sciences, as well as in science, technology, and medicine (STM), that document their developers' selection criteria. Interviewees who are directly involved in "collecting" free Web resources were asked to describe their experiences hiring and training staff, establishing and applying selection and cataloging criteria, identifying resources, and developing end-user interfaces.
2. *Web Browsers*. Using basic and advanced search strategies, the phrase "selection criteria" was used to locate Web sites where these terms were used and where criteria for selecting Web-based resources were discussed. This methodology was the least valuable of those used in researching this report, but it did confirm that little information on this topic is easily retrievable through browsers. Web browsers also permitted the identification of a number of highly developed subject gateways, the policies and practices of which proved very relevant to this report.
3. *Gateways*. More than 50 subject gateways, ranging from small to large, were reviewed in an effort to locate and evaluate their selection criteria, staffing patterns, and cataloging and related practices. Reviewing these sites and communicating with their staff was the most useful of all the forms of data gathering used for this report, and it yielded by far the most valuable findings. Creators of European and Australian sites have refined and articulated selection and related criteria to a far greater degree than have their North American colleagues. The selection criteria and documented practices of gateways such as the Internet Scout Report, however, approach or equal the high standards set by gateway creators in the Nordic countries, the Netherlands, Germany, the United Kingdom, and Australia.

4. *English Language*. Because this report was written for professionals involved in developing digital collections in North American libraries, only sites that provide documentation in English were evaluated. Owing to the extent to which the English language dominates the Web, limiting the report to English-language documents did not restrict the scope of the investigations upon which it is based. On the contrary, many of the most promising practices discussed in this report are for gateway projects in non–English-speaking countries and in which the documentation and user interface are provided in English and at least one other language.

1.2. Content

This report focuses on issues pertaining to the development of sustainable collections of free Web resources. The findings that emerged from the research done for this report document that developing and managing collections of free Web resources have wide-ranging, long-term implications for human resources, organizational issues, and fiscal matters that extend well beyond the circle of individuals responsible for selecting these resources. What one selects and how one wishes to deliver it to the end user frequently have significant implications for workloads and priorities in traditional cataloging and metadata departments, technical support units, and user-services programs. Free Web resources are challenging how libraries have traditionally processed and added value to print and analog publications. Many of the practices and policies developed by the projects reviewed for this paper shed light on how to address and manage the myriad access, staff development, user training, and cost issues associated with collecting free Web resources. Thus, in addition to the specific collection-development policy issues of collection scope, selection criteria, and resource discovery, a discussion of the broader issues pertaining to managing collections of free Web resources is included. Among these issues are the following:

a. *Added Value or Cataloging*. What levels of description, ranging from free-text notes to MARC records with traditional subject headings and metadata, are necessary or adequate to provide efficient access to free Web resources?
b. *Access and Archiving*. Options for integrating free Web resources into collections and services must be explored. For example, do OPACs provide sufficient and appropriate access? Are separate files required, or are new Web pages with a series of links adequate to ensure users will find these resources once they are selected? Once the resources are identified, how can extended availability be assured, given the ephemeral nature of Web-based content?
c. *User Support*. How do users learn of free Web resources after they have been selected? What are the end users' support needs, and what are the implications of these needs for staffing and services?

d. *Human Resources*. What expertise is needed to select wisely and efficiently? The qualifications for staff who select Web resources are not as clearly defined as they are for staff who select print-based and analog resources for research library collections. How does one incorporate selection of free Web sites, a new job responsibility, into existing workloads? What skills and expertise best qualify an individual to be involved in the selection, cataloging, and technical support processes?
e. *Workflow and Organizational Issues*. Making free Web resources available to users directly has an effect on multiple work units and workflows. Decisions pertaining to Web resource selection, processing, and delivery have specific implications for preexisting workflows and priorities. In building sustainable collections of free Web resources, planners must recognize that there are connections between decisions that relate to Web resources and decisions that relate to analog collections.
f. *Fiscal Implications*. Free Web resources are not unlike gift collections: their acquisition has direct budgetary and workload implications. Identifying, selecting, and making available free Web resources incur costs to the library. These costs are not trivial, and they can ultimately influence both the selection and access delivery processes. Identifying the implications of free Web content and at what cost it will be managed and made accessible to end users influences planning, outcomes, and costs throughout the library.

1.3. Terminology

The research performed for this report revealed that not all individuals recognized for their Web resource expertise use technical terms or jargon uniformly. In some cases, variant terms were used synonymously; in others, experts used the same terminology quite differently, or at least with a variant nuance. To prevent ambiguity and to ensure consistent usage, it was deemed appropriate to define several terms. In reviewing these definitions, readers should bear in mind that much of the literature in which these terms are used most consistently has been prepared in the European Union, where British usage is more common. The definitions presented here, however, do not vary significantly from North American usage.

Free. Describes Web-based resources for which no compensation is required by the creator or host site in order to have full access to the content.

Collection. Two or more related items, the relationship of which is determined by the scope of the collection, as defined by the staff responsible for building that collection.

Gateway (i.e., subject or information gateway and synonymous with subject guide, subject page, megasite, virtual library, clearinghouse, and, increasingly, portal). The Internet Detective defines a "subject gateway" as differing from a search engine in that the content of a gateway has been selected through some form of human input, normally a critical evaluation by an information professional or subject expert. IMesh defines a subject gateway as "a web site

that provides searchable and browsable access to online resources focused around a specific subject. Subject gateway resource descriptions are usually created manually. For this reason, they are usually superior to those available from a conventional web search engine" (IMesh Toolkit 2000).

Andy Powell provides a detailed set of definitions for the Resource Development Network (Powell 2000).

2. WHY SELECT FREE THIRD-PARTY WEB SITES?

Today, researchers in all disciplines are creating scholarly Web-based resources, and students and faculty regularly browse the Internet in search of sites related to their interests. Some create profiles describing their needs so that search engines or harvesters relying on user-friendly language-retrieval systems can locate sites matching a given profile and notify the user. What is the role for librarians, other subject experts, and information specialists in developing relatively labor-intensive and costly handcrafted collections (i.e., subject guides, catalogs, gateways) of otherwise freely accessible Web sites?

Many free Web-based resources are merely collections of links to other sites or combinations of links and scanned files of texts, images, or sound and video. A growing number contain information not readily available elsewhere. Once found, many of these sites are easily viewed and downloaded. More sophisticated sites may require specialized or domain-specific tools, technologies, or guidance to view or use certain types of content (e.g., geospatial, scientific, or statistical information). These sites are difficult or impossible to find, even by scholars and information specialists. Search results can be so overwhelming that the user cannot be expected to evaluate them in a reasonable length of time. Moreover, organization within the Web presents most scholarly users with a technological labyrinth. Results may be far removed from, or totally unrelated to, the desired findings. Finally, the artificial intelligence technologies employed by the major Web discovery tools are insufficient to retrieve and adequately evaluate scholarly content.

Many of the inadequacies of the Web are due to the fact that the Web is still best described as "quantity without quality." A more accurate statement would be that the Web reflects the full spectrum of quality, ranging from minimal or null to highly authoritative. Searching the Web on any topic will retrieve all information pertinent to one's query, but there will be no qualitative evaluation or filtering of the content. Even the relevance rankings offered by some browsers do not reveal the quality of the information retrieved; they reveal only the frequency of terms stated in the query. Thus, a recent google.com search for "Mozart" retrieved more than 800,000 sites. They offered information not only about the composer and his works but also about composers influenced by Mozart, opera houses and companies, music processors, programming software, greeting cards, museums, the Salzburg airport, Austrian bonbons, and even person-

al pages in which the name Mozart appeared. Refining the search to "Wolfgang Amadeus Mozart" reduced the number of hits by a factor of 10, to a still-unwieldy 78,000+ sites. "Mozart biography" retrieved 322, and then "about 392" sites; "biography of Mozart" yielded "about 437" sites.

In cases such as these, relevancy ranking can significantly facilitate resource discovery. Search engines that display results according to relevance do so on the basis of the following factors:

- search term frequency in the document
- inverse document frequency of the search term (i.e., the lower the frequency of the term in the entire database, the higher the relevance)
- length (density) of the document
- date of the document
- occurrence of the term in the document title or only within the document
- proximity of two or more terms (the greater the distance between terms, the lower the relevance)
- link popularity (i.e., the frequency with which other Web resources link to the document); similar to citation frequency measures used to evaluate the importance of a print resource or its author

Unfortunately, relevancy rankings do not ensure precise retrieval. They do not provide the higher education community with an evaluation of the intellectual level of content, nor do they provide the levels of organization and selectivity students and faculty require to take full advantage of free Web-based information. Only when sites have been reviewed, evaluated, selected, and cataloged will users be spared the ambiguities resulting from the randomness and "quantity without quality" of Web search results. Until technological advances permit rapid and highly precise retrieval of only that free Web-based content that is both authoritative and useful in the context of higher education and research, expert human intervention will be needed.

And therein lies the role of librarians and other subject specialists. The value-added services that libraries have traditionally provided for print formats need to be applied to free Web-based resources as well. The selection and cataloging functions of a physical library assure users that titles found there have met predetermined quality criteria and have been described in a manner that facilitates their identification and retrieval. Moreover, the cataloging process provides the authoritative and consistent grouping of related materials so vital to browsing and to the winnowing-and-sifting process that characterizes learning and research in an academic setting. As Sarah Thomas states, "Libraries can add value by promoting filtering and ranking which would prefer [Web-based] resources . . . that meet a set of established criteria, such as having a strong likelihood of authenticity, accuracy, or endorsement by others of standing" (Thomas 2000).

Another traditional library service that is absent in the Web environment is storage or archiving. Once print resources have been selected and integrated into a research collection, most of them will be retained indefinitely, or at least until their physical condition requires withdrawal or their content has been superseded to such a degree as to warrant replacement. This is not the case with Web-based information. Web content is not durable; there are no archival guarantees. Information available at any given moment can move or cease to exist without warning. Were libraries to select and archive Web resources in much the same way as they have collected print formats, users could be assured indefinite (i.e., archival) retrieval and delivery of what is otherwise an ephemeral format. How best and at what cost free sites should be archived and by whom are questions currently under discussion. (See Section 5: Data Management: Collection Maintenance, Management, and Preservation for further discussion of issues pertaining to archiving Web-based information.)

Librarians and other subject specialists have myriad opportunities to select and organize Web-based information. The current realities of the Web validate the need for human intervention in facilitating access to Web-based resources within the higher education context. By selecting a subset of resources that meet predetermined criteria and by facilitating access to them, librarians impose a quality structure on those resources. Users can be assured that sites found in such collections—selected, evaluated, and described by persons with recognized subject expertise, and made available through a catalog that provides a single interface to an integrated collection—are of related quality. Users can also be assured that access to these sites will be stable, and that resource discovery will be tailored to the characteristics of the collection and its content, rather than to the features of a specific search engine.

For a description of the role of librarians and subject experts as performed in an applied setting, see Lagoze and Fielding 1998.

3. IDENTIFICATION, EVALUATION, AND SELECTION

3.1. The Need for Web-Specific Selection Criteria

During the second half of the twentieth century, the library literature devoted much space to collection development and management theory and practice as they pertain to print and analog formats. Library school curricula and professional development workshops gave considerable attention to these topics as well as to selection practices and polices. Learning the component parts of collection policy statements and drafting sample statements were basic requirements in library school programs, and for many years they provided the focus for professional development and continuing education forums. In addition, library administrators have traditionally expected selectors to apply these theories and principles in formal policy statements. Although written policy statements frequently lie dormant

for extended periods, they steer day-to-day selection decisions and determine the nature and quality of collections over time. Formally articulated collection-management policy statements are crucial to collection building in that they document the intended scope of a collection (i.e., what it excludes as well as what it includes) and explain historical strengths and weaknesses within a defined context.

This long-established practice of documenting collection policies for print and analog formats notwithstanding, there was at the close of the 1990s a paucity of documentation describing the practices followed in North American academic libraries when selecting free Web resources. Many subject specialists developing guides to free Web resources do not recognize the need to articulate their collection policies, as even a cursory survey of the Web will reveal. Online policy statements are rare; written documentation describing the goals and objectives in creating and maintaining these guides are rarer still.

This absence of formal, well-articulated, and commonly accepted criteria or guidelines for selecting free Web resources is due in part to the way in which such resources have thus far been acquired and added to academic library collections. By and large, administrators and subject specialists alike have perceived the selection of digital resources to be an extension of the print selection process. Staff charged with selecting print resources are the primary selectors for all or most digital formats, whether or not those resources are free. Blurring distinctions further is that information resources produced by and for academic and research communities have many similarities, whether these resources are print-based, analog, or digital. Scholarly resources can be textual, image-based, numeric, or geospatial. Their creators can be personal, corporate, governmental, nongovernmental, or even anonymous. Their content can be fact or fiction, original works of art or music, interpretive performances, or critical studies. Despite these differences, their single intended audience remains the academic community. Given this scenario, the need to articulate new or additional selection criteria for free Web-based resources is not immediately apparent. Indeed, many of the evaluative criteria that apply to print-based publications also pertain, to varying degrees, in the Web environment. The most obvious of these criteria are *content quality* and *programmatic need.*

As appropriate as these and other traditional selection criteria for print and analog formats may be, they quickly prove insufficient when selecting Web-based resources. Many print-based criteria lend themselves only partially to the evaluation of Web-delivered content; other criteria are not applicable at all. Web resources exhibit qualities that do not exist in print and analog formats and for this reason require additional evaluative measures. To evaluate their content, format, and dependence on technology and to describe their value and applicability to higher education and research, new criteria are needed.

Free third-party Web resources are particularly challenging. Their availability does not bear the imprimatur or *nihil obstat* that the traditional scholarly vetting process confers on the printed word. Frequently, the authority of those who have created these resources

is not immediately discernible, and the veracity of their content cannot be surmised from the names and reputations of scholarly publishing houses. The concept and role of publisher in the case of free resources are oftentimes vague. Is the publisher the site hosting the file where the content is found? Or is it the person or body that developed the site? Further, most of the traditional services associated with marketing and distributing print resources are lacking. Vendors do not facilitate the acquisition of Web resources through any type of notification service, blanket order, or approval plan. National bibliographies do not record their existence. Rarely if ever do unsolicited mailings or the traditional print or electronic blurb announce their "publication." Only a small number, far fewer than one percent by anyone's measure, receive a formal review. Simply stated, the basic selection guidelines and principles of how to identify, evaluate, and acquire print-based and analog library materials—principles that are articulated so clearly in the collection-development canons—do not pertain in the environment of the Web. The successful inclusion of Web resources in academic library collections requires new or, at a minimum, additional selection criteria.

Fortunately, the number of clearly articulated, high-quality selection criteria for free Web resources is growing. The American Library Association (ALA) has promulgated basic guidelines for selecting Web sites suitable for public libraries, reference services, and the fields of education and business (2000, 1999, 1997). Some individuals have published well-written articles on general principles pertaining to selection of free Web resources (McGeachin 1998, Fedunok 2000, Sweetland 2000). A valuable collection of articles on major projects to harvest and organize Web-based content is *The Amazing Internet Challenge: How Leading Projects Use Library Skills to Organize the Web* (Wells, Calcari, and Koplow 1999*).* The extent to which staff have applied or adapted these guidelines in other projects remains largely undocumented.

Nevertheless, Web selection criteria are only rarely found in professional journals and other print formats. Instead, they are customarily found as Web-based documents, imbedded in scope statements and collection policies of the increasing number of subject gateways and in the background papers on resource discovery systems. For example, well-delineated and highly articulated criteria can be found at Web sites developed by the Internet Scout Report, the Research Discovery Network in Great Britain, sites in the Netherlands (e.g., DutchESS), and at various Scandinavian sites, such as those developed at the Kuopio University Library in Finland, all of which are available in English. The Internet Detective and the DESIRE Handbook are also valuable.

What follows is both a synthesis and an amalgamation of criteria and practices for evaluating free Web sites described at these and related sites in North America, Europe, and Australia. These criteria cover the full range of issues associated with evaluating free sites: provenance, authorship, content, design, user support, standards and technologies, navigation, and system integrity. The criteria have

been selected for inclusion in this paper because of their universality. They apply equally to scientific, technical, medical, social science, and humanities Internet resources. Their application in existing projects has led to the creation of consistent and coherent collections of "high-quality" content. The *DESIRE Information Gateways Handbook* uses service-driven definitions to describe "high quality." Thus, "quality" depends on users' needs, and a "high-quality" site is one that meets their needs because it is relevant for their purposes. Further, quality is determined on the basis of inherent features of a resource. Thus, high quality can be ascertained only if a skilled and knowledgeable human evaluates both user needs and the inherent features of the Web resource. Building on this premise, Kuopio University Library staff expressly state that "Information content should comply with the needs of the frame organization" (Kuopio University Library Group 1996). That is to say, the resources selected should support the teaching, research, and information services of the library selecting them for their collection.

3.2. Developing Selection Policies

The need to define collection parameters and to articulate content criteria is equally important in the print and digital contexts. Development and maintenance of consistent and coherent collections of high-quality print or digital resources can succeed only in an environment in which value judgments are made on the basis of previously defined and agreed-upon collection policies and selection criteria. Clearly articulated policies allow staff, over time, to select content that is consistent with their institution's mission and long-term goals. Formal policy statements afford users an opportunity to understand and evaluate the rationale underlying a collection and to determine its ultimate value in relation to their needs and expectations. How users conceptualize the nature and features of a collection significantly influences how they perceive that collection will serve their needs.

The *DESIRE Information Gateways Handbook* notes the following advantages of developing formal policies specific to free Web resources and making them available online to both staff and users. Such policies:

1. help users appreciate that the service is selective and quality controlled;
2. help users understand the level of quality of information they will find when using the service;
3. help staff be consistent in their selection and to maintain the quality of the collection;
4. help train new staff; and
5. ensure consistency in collections that are developed by a distributed team.

According to this list, formal policies are equally valuable for staff and users. Similarly, quality and consistency receive equal emphasis; this underscores the need for a formal policy that staff members apply and users understand. Only when selection practices are broadly understood and consistently applied can quality be defined and identified. Through the consistent application of accepted practices, the integrity of the collection (i.e., the value of the individual pieces and of the collection as a whole) will be assured and consistent.

As these five points demonstrate, the fundamental principles and components found in policy statements pertaining to collections of print and analog formats apply to a certain extent to free Web resources. Understanding and acknowledging this fact will facilitate developing and articulating policies suited for free Web resources. This understanding will also assist staff in recognizing that Web resources are a continuum in the history of scholarly communication, and that their intellectual content is not inherently different from that of other formats. Nonetheless, while much of the rationale for developing collection policies in the print and analog environment applies to developing collections of free Web resources, evaluating the full essence and nature of free Web resources requires additional evaluative criteria—criteria that evaluate the unique aspects of information contained in and delivered through free sites.

3.3. Collection Scope Statement

A well-developed collection policy includes a *scope statement* as well as *selection criteria*. The scope statement is the first filter in the resource evaluation process. It defines the parameters of the collection in broad terms by describing what is included in and excluded from the collection. The justifications for these inclusions and exclusions should be clearly stated and based on the needs of the intended users. At a minimum, the scope statement outlines the subjects included in the collection. It clarifies the acceptable sources for information (e.g., academic, government, commercial, nonprofit, personal resources) and the acceptable level(s) of difficulty (i.e., suitable for use in higher education settings, scholarly, kindergarten through twelfth grade, or popular) that will be considered. The scope statement also describes the resource types presumed to be relevant to the subject and the primary audience. The following list of resources included in the Humbul Humanities Hub can serve as a guide to defining collection scope:

- primary/original sources in electronic format (i.e., information mounted on the same server as the site and created or produced by the owners of the site)
- secondary sources, whether published solely on the Web or as surrogates for printed editions (i.e., "third-party information," located and created at another site and made available through a link

that takes the user to the other site; secondary information may be primary information at its host site)
- research projects and reports
- bibliographies and bibliographic databases
- electronic journals
- e-mail lists where online archives exist
- academic department Web pages
- professional association Web pages
- resource directories

If resources that one might expect to find are excluded on the basis of access restrictions (e.g., technological barriers, cost, or registration) or of content, the scope statement should list these exclusions and give the rationale for not providing access to them.

Other aspects of scope concern language and geographic parameters. Although often overlooked because of the predominance of English-language resources on the Web, the increasing number of bilingual and multilingual high-quality sites calls for greater consideration and clarification of the language parameters imposed on a collection. The growing number of scientific sites, especially mathematics sites, where numbers, signs, symbols, and formulas are not language-specific demonstrates the importance of evaluating narrow or highly restrictive meta-language parameters before they are put in place. Similarly, geographic parameters need to be carefully reviewed and clearly stated. The Web is one of the primary forces reducing, if not erasing, geographic barriers to communication and access to information. The decision to limit the scope of a collection to resources developed in a particular country or in a particular language may be appropriate, but the justifications for that decision need to be stated.

RECOMMENDED EXAMPLES OF SCOPE STATEMENTS

DutchESS. Dutch Electronic Subject Service. DutchESS Scope Policy. Available at http://www.kb.nl/dutchess/manual/scope_eng.html.

Humbul Humanities Hub. Available at http://www.humbul.ac.uk/about/colldev2.html.

Internet Detective. Creating a Scope Policy for Your Service. Available at http://www.netskills.ac.uk/TonicNG/content/detective/56.html.

Jennings, Simon. 2000. RDN Collections Development Framework. Version 1.1 (May). Available at http://www.rdn.ac.uk/publications/policy.html.

Library of Congress. BEOline: Selection Criteria for Resources to be Included on the BEOnline+ Project. Available at http://lcweb.loc.gov/rr/business/beonline/beonsel.html.

SOSIG Social Science Information Gateway. Training Materials and Support. Available at http://www.sosig.ac.uk/about_us/user_support.html.

3.4. Selection Criteria

Although resources that do not fall within the parameters of a collection scope statement should be rejected without further review, not all resources that do fall within the scope of a collection necessarily warrant inclusion. Further review and evaluation using established selection criteria assure that a collection will include only those materials that best meet users' needs and expectations (i.e., those that meet the definition of "high quality"). Responsibility for setting these selection criteria varies from one gateway to another. Some gateways openly state that "no [selection] criteria were established" (e.g., ALA Machine-Assisted Reference Section [MARS] (1999)). Others provide brief selection criteria statements (e.g., Best of the Best Business Web Sites [BRASS]). An increasing number of specially funded projects are built on highly articulated selection policies (e.g., Dutch Electronic Subject Service [DutchESS]; Engineering Electronic Library [EEL], Sweden; and Internet Scout Report).

The selection criteria described in the selection policies and collection statements reviewed for this paper can be categorized into four groups: context, content, form/interface, and technical criteria. Each of the four groups comprises multiple criteria. No one criterion will adequately evaluate a site, and frequently several criteria overlap and intertwine—a process that blurs their definitions yet ultimately reveals more effectively the quality of the site under evaluation. A review of the collections using the following criteria did not reveal that any of them are more important than any other; moreover, no stated or implied priority order for applying these criteria can be found. Therefore, the order in which criteria and their categories are presented here should not be construed to imply any hierarchy or relative importance. Rather, they appear in an order that might facilitate the culling process. Their inclusion is based on their repeated occurrence among selection criteria used and described by various projects and programs committed to the development of high-quality collections of free Web resources.

3.4.1. Context Criteria

"Context" applies to the origin (provenance) of a site and its content, as well as to the suitability of a new resource to an existing collection. How a site meshes with and enhances existing or anticipated content within a collection will determine the quality of the collection over time.

3.4.1.1. Provenance. The origin or source of a site reveals and confirms much about its value and provides important information about its overall quality. Parsing a site's URL often yields sufficient information to judge the affiliation of its creator and the reliability of its server.

3.4.1.2. Relationship to Other Resources. Like print resources, free Web resources can stand alone or in aggregate with other resources. As with any collection, however, the value of a collection of Web sites lies in the integrity of its individual components. The greater the degree to which each site (component) within a collection relates to and enhances others within the same collection, the greater its value of the collection as a whole. Evaluations of free Web sites have shown that many of them contain links to the same sites. This redundant content will increase the size, but not the quality, of the collection. Each site added to a collection must be viewed as an integral part of a larger mosaic. Redundant, superfluous, unrelated, or poorly suited pieces will not enhance the collection; they will only encumber it and ultimately discourage or confuse users.

3.4.2. Content Criteria

Content is arguably the most important criterion used when selecting resources, and it must apply equally to all sites. The term applies to the information contained in a resource, and it is used both qualitatively and quantitatively. A high-quality site with little content may or may not prove to be a useful site; however, the content of a low-quality site that has high quantity requires close scrutiny, for if little else exists on a specific subject, its mere availability may prove useful. Content should be evaluated on the basis of multiple factors and ultimately judged on the degree to which it supports the purpose of the collection to which it will be added.

3.4.2.1. Validity. Editors, reviewers, and publishers carefully vet the content of printed scholarly resources before they are published. This is not true in the case of free Web resources. Anyone with certain basic skills and access to server space can make information accessible to anyone who happens upon it. Web-based information that is flawed, whether intentionally or inadvertently, is not always easily discernible from information that has been carefully and thoroughly verified. Especially when combined with high-quality Web features such as attractive graphics and slick navigational functions, misinformation, false claims, and factual errors may not be self-evident. For example, an essay by I. Newton on the theory of gravity or by W. von Braun on rocketry may not contain the ideas of the great scientists these names imply, but rather the work of Ian Newton and Wesley von Braun, 12-year-olds hard at work on their middle school science projects.

The validity of free Web resources should be measured on the basis of several factors. Is the source of the content clearly identified? Does the URL support the claims of the "author's" affiliation and credentials? Is contact information readily available, and, if so, what does it reveal about the person or team responsible for the content? Is the information available in print? If so, does the print version imply or state a critical review or validation of the information contained therein? To what extent has the Web resource been vetted by a third party; has it passed through any kind of "quality filter"?

Is there evidence of or description of the extent and nature of quality checks, review, or evaluation of the content? Is the source of the information adequately described? Is the research process leading to the creation of the content described? Are references cited? Is a bibliography or webography provided?

3.4.2.2. Accuracy. Whereas validity measures the degree of objective truth, accuracy is a measure of the degree of correctness in the details. A resource may exhibit validity but lack a high decree of accuracy. Not all books that have successfully passed through the vetting process because of the validity of their content are as highly reviewed when evaluated for the accuracy of their supporting details. To some degree, the same is true of Web site content, which may exhibit inaccuracies caused by errors in data entry or by intentional biases of the person or team who developed the content. Even errors in spelling and keying are important indicators of the accuracy of the content. A high error rate in one aspect of a site (e.g., keyboarding) is often a good indicator of other potential flaws. Many of the measures for judging validity also apply to evaluating accuracy. Judging accuracy is difficult. It requires subject expertise and can prove to be highly subjective.

3.4.2.3. Authority. The authority of a site is dependent on the reputation, expertise, and credentials of the person(s) responsible for creating the site and providing access to it (in traditional terms, the author and publisher). The experience and credentials of individuals who create free Web sites vary radically. Authors may be enthusiastic hobbyists or recognized authorities who have devoted years of study to the topic. Does the site provide biographical information or resumes about the author? Knowing whether the site content is attributable to Sir Isaac Newton or to the hypothetical Ian Newton referred to earlier may provide an important measure of the site's authority. Similarly, the location of the server may prove valuable in evaluating authority. Servers maintained by colleges, universities, museums, scholarly societies, or professional associations imply high levels of expertise in the development of site content. URLs containing a tilde (~), even if on a server maintained by such institutions or organizations, require closer evaluation, because a tilde normally signifies a personal page that is merely maintained at that location. Other indicators of authority are postal and e-mail addresses that can support authors' claims regarding their credentials and affiliations. The primary measures of authority are

- the creator of the site
- the creator's reputation
- where the server is located
- how many other sites link to it

Sites that do not identify authorship require particularly close evaluation through the application of other content criteria. Anonymity is normally not associated with high-quality content.

3.4.2.4. Uniqueness. Uniqueness is a measure of the amount of primary (i.e., original) information contained in a site. It also is a measure of the extent to which information at a site under review is duplicated in sites that have already been selected for the collection. Some degree of content redundancy is acceptable; one common characteristic of free Web resources is that their content duplicates that of other sites. The context and manner in which duplicate information is presented at other sites may argue in favor of selecting sites that contain information currently in a collection. Redundancy, however, must be evaluated carefully. The nature of digital information and its accessibility by multiple simultaneous users nullify all legitimate arguments for duplicate copies of information in print and analog collections. Users searching a specific collection will not be well served if their search yields multiple hits, all of which point to the same information and accessible through multiple sites. Sites that contain primary information not available through other sites, depending on how the site measures up against other criteria, will undoubtedly prove to be of greater value than will sites containing secondary information—unless the secondary information offers significant added value. "About this site" links can prove highly beneficial in evaluating uniqueness. An evaluation of the URLs to which the site links will reveal whether they point to information at the site (primary information) or to external sites that have probably been created by someone else. Links to external sites can be desirable, particularly if they offer significant added value.

3.4.2.5. Completeness. Completeness does not mean comprehensive coverage of information or exhaustive treatment of a topic. Rather, it refers to the availability of content at a particular site. The phrase "under construction" signals that the creators of the site are eager to inform others of their work but that the content or site design remains incomplete. However, the absence of the phrase "under construction," similar wording, or icons conveying that message does not assure that the content is indeed complete. A site is incomplete if it points to non-networked resources or to print resources for the full "edition" of the "work" or if links are grayed out. Some links may not be grayed out but point only to empty files. The content may fail to support the purpose of the site; this is a serious indication that the site is not complete. The extent to which the content of a site is incomplete and the length of time it remains incomplete affect the quality of the site and the quality of any site or collection that links to it.

3.4.2.6. Coverage. Coverage refers to the depth to which a subject is treated. The term is used both qualitatively and quantitatively. How adequately the site treats the topic and to what degree (i.e., depth) it treats it determine the quality of coverage. Is the topic sufficiently represented, or are the various aspects of the topic treated superficially or not at all? Is there primary information? If little or none, how successfully does the site provide linkage to primary content maintained at other sites? What are the gaps in coverage? Even the

finest print collections have lacunae, but there is a limit to the extent and nature of gaps, in both print and digital collections, beyond which the limited coverage affects the quality of a specific item or the collection generally. Links that are present, but broken, adversely affect coverage. Indexes, contents pages, and bibliographies facilitate the evaluation of the coverage; however, sites that are limited only to links or that consist only of enumerative bibliographies may prove inadequate if the bibliographic information is not supplemented with abstracts or links to the full texts. On the other hand, if the site provides the only bibliography on the subject or provides more substantive or current content than do other bibliographies, it may be an important site solely on the basis of its bibliographic coverage. At this time in the history of the Internet, users should not expect to find that coverage of a topic provided by a free Web site necessarily equals that of many print resources. Internet publishing, though maturing rapidly, is in its incunabulum phase.

3.4.2.7. Currency. Currency pertains to the degree to which the site is up-to-date; i.e., the extent to which it presents prevailing opinions, ideas, concepts, scientific findings, theories, and practices relating to the subject. Policy statements, for example, on the selection of "electronic resources" written in the early 1990s and concerned primarily with the evaluation and acquisition of information on CD-ROMs would appropriately receive a low currency score today because the digital formats currently available were unforeseen at the time the site was created and criteria for their evaluation had not yet been articulated. Currency, therefore, refers to content that describes the current thinking and the context in which higher education finds itself currently. Just as a library collection of print materials must reflect the latest research, so too should high-quality free Web site content. (See also a discussion of Information Integrity in section 3.4.4.1.)

3.4.2.8. Audience. Audience is what traditional print selection criteria describe as "intellectual level of content." That is, for whom is the information intended? Once the intended audience is discerned, the person selecting the site should ascertain how well the site meets the needs and interests of that group. A high-quality site on organ transplants that is created for middle-school students will not serve the needs of university students. The site itself could be outstanding, but if it fails to meet the needs of its intended users, it is of low quality.

3.4.3. Form/Use Feature (Accessibility) Criteria

Form criteria are features that determine how the content is presented and how accessible it is. They include, for example, organization and user support. How readily a user can access content at a site, display it on a monitor, and download or print content depends on site design, user aids, and software applications. Unlike print resources, Web sites are not uniform in structure. Their access and navigation are neither uniform nor based on centuries of tradition-bound organizational concepts. Web resources lack the physical structure and

content sequence (e.g., title page, foreword, introduction, table of contents, contents, and indexes) that characterize how printed matter is organized and that facilitate access.

Well-developed and organized digital content is far more than a series of pages to be read sequentially or accessed by page number or by specific indexing terms. Rather, it is multifunctional, allowing dynamic searching and downloading of content. Display, however, is a two-dimensional image that lacks the physical cues that permit or enhance use. Links and icons facilitate access to the content and permit the user to comprehend the purpose and content of a resource and use it effectively. Because these features may differ in quality, form (use features) must be as carefully evaluated as is content when selecting Web resources.

3.4.3.1. Composition and Site Organization. Because of the peer-review and editorial processes associated with print publications, an analysis of how information in a print format is organized, structured, and arranged requires only minimal evaluation during the selection process. Editors and reviewers contribute greatly to the final form of a printed work; spelling errors, bad grammar, poor style, and poor organization of content are removed in the vetting, revision, and editorial processes. Web resource creation does not consistently include equally rigorous evaluation and revision as does information in print format; in fact, the most extensive evaluation of composition and organization may occur after completion and then only by end-users. How well a site is composed and organized will determine how accessible users find the content of any Web resource. Thus, the selection process must include an evaluation of whether Web site content is logically or consistently organized or even divided into logical, manageable components that meet the needs of the intended users. Further, evaluators should take into consideration whether design enhances or hinders accessibility. A site may prove be "overly designed" if aesthetic features (e.g., wallpaper, graphics) require an unnecessary use of plug-ins or hinder access because they make images slower to load. Not all users have state-of-the-art computers or high-speed modem connections. Ease of access is relative and should weigh heavily in site evaluation (Wells, Calcari, and Koplow 1999, 210). For a good discussion of issues pertaining to composition and site organization, see "Training Materials and Support," on the Social Science Information Gateway (SOSIG) Web site.

3.4.3.2. Navigational Features. In addition to software applications (such as special viewers), layout, design, search functions, and user aids can facilitate or impede navigation of a site. Each of these features requires a separate evaluation, but it is their successful combination that produces a high-quality site. One measure of navigational quality is browsability. How easily can the user browse the content? Are logically organized subsets of related information grouped so that manageable amounts of data or information can be browsed? Browsability is directly related to the value of a search

function within a resource. Does one exist, and is it adequately designed and described? The availability of keyword search functions combined with Boolean operators should be expected.

The value of search results, however, depends on the quality of indexing. Whereas search functions pertain to how one searches; indexing pertains to choice of terms and to the depth to which terms are linked to content. Indexing and searching should, therefore, be evaluated separately. Superior search functions provide little added value if poor quality indexing standards or practices have been applied. Once a search has located information within the resource, how conveniently can the user print or download relevant content? For example, can a user print or download a single file of documents that is composed of a series of separate pages? Some site criteria suggest that substantive content should be no more than three clicks away, and that all links should be unambiguous and make it clear to the user where links are leading (see SOSIG). Some navigational features (e.g., tables of contents and indexes linked to content and buttons or icons directing users "home," "forward," and "back") are quite simple.

3.4.3.3. Recognized Standards and Appropriate Technologies. Closely related to composition and navigational features are the standards and technologies used in developing a site, for they determine the degree to which a user may access and use Web-based content. Today, accessibility weighs heavily in decisions pertaining to both technology and design, and closer attention is paid to developing and adhering to minimum design standards that will assure access to the largest number of users possible. The U.S. Department of Education Office of Civil Rights has defined these standards as those that ensure "timeliness of delivery, accuracy of the translation, and provision in a manner and medium appropriate to the significance of the message and the abilities of the individual with the disability" (Waddell 1998). Currently, the implementation of the Americans with Disabilities Act is receiving close review in higher education circles, for accessibility to the Web obtains whether one wishes to assure access by individuals with disabilities, distance learners, or users with slow modems and low bandwidths. The World Wide Web Consortium (W3C) has issued Web Content Accessibility Guidelines 1.0 (Chisholm, Vanderheiden, and Jacobs 1999).

Before selecting free Web sites, one must consider the range of these standards and technologies and determine which are most critical for prospective users. Two fundamental questions are whether sites function in generally available environments or whether special extensions are required. All users expect content to load quickly and access to be rapid and available at any time. The Internet Detective provides important introductory guidelines for evaluating accessibility and for determining how it is affected by the standards and technologies that designers employ. For example, does the resource use proprietary extensions to HTML that some browsers cannot recognize? Can the information be accessed if the user has keyboard-

only navigation capabilities? The presence of metadata, alternative text when images are switched off, text captioning for audio material, and links and downloading instructions to any software that is required to use the resource are key to evaluating the extent to which currently accepted or recommended standards and technology have been used in developing a site. Simply stated, are the technologies used appropriate for the content and the intended audience? One basic question that should not be overlooked is whether a site offers users more than they might find in a print resource containing the same information. For further details on this subject, see the Internet Detective.

3.4.3.4. User Support. User support relates to content (e.g., scope statements and collection policies) and technological applications. A well-developed Web resource provides information that answers users' questions as they arise. Frequently, this information is provided through frequently asked questions (FAQs), which anticipate the most commonly requested or needed information about using the site.

FAQs, however, provide only static support. In some cases, this is sufficient. For example, if a particular viewer or audio software is necessary or preferable to provide full access to content, a simple statement to that effect may prove adequate for the user to proceed. When, however, interactive support would be more beneficial to the user, help screens or contact information should be easy to locate and user-friendly. "Contact Us," "Help Desk," or similar links (e.g., e-mail, postal address, phone numbers, office hours, individuals' names) should be included in the design and be easily identifiable.

3.4.3.5. Terms and Conditions. Access to some free Web sites requires that users first agree to specific conditions. These may range from simple registration to user fees or agreements to download information only under specific conditions. A great deal of content delivered through the Web that supports higher education and research is accessible without any restrictions. Where restrictions do exist, one must evaluate whether they are appropriate. If registration is required, why is it necessary? How will user information be retained and used or shared with third parties? Registration information may be gathered to assist site developers in monitoring what and how users access content in order to improve content coverage and access over time. It may also be gathered for reasons that are not apparent or that go unstated. Where user fees are applicable, are they appropriate? If fees are required to access a site, does the service point only to resources that are themselves free? If so, does added value at the site warrant the required cost to users? Terms and conditions of use per se may not be inappropriate, but their validity should be confirmed before a site is added to a gateway or OPAC.

3.4.3.6. Rights Legitimacy. Related to terms and conditions of use is the question of whether those responsible for the content and access to the site have the right to place restrictions on the content at that site. It is relatively easy to claim rights to information, but such claims may be unfounded and, therefore, not legitimate. Proving the legitimacy of such claims is difficult and time-consuming.

3.4.4. Process or Technical Criteria

Among the most frequently cited characteristics of Web resources is volatility. Links within a site may abruptly cease to function; an entire site may disappear without warning. Free Web resources come with no guarantees that their content or accessibility will remain constant or improve over time. In extreme cases, an entire resource may cease to exist or change its URL without directing users how to access the same resource at a new URL. Process criteria are those technical features that measure the integrity of a site and the availability of content reported to be provided by the site. They are the functions and features that permit the end user to access the content selected and described by the provider (i.e., author or creator).

3.4.4.1. Information Integrity. Information integrity, or maintenance, pertains to the intellectual content offered by a site and is measured by how successfully the value of the content remains current or improves over time. Determining the value of content at any site requires knowing when a site was created, at what frequency updates and revisions are scheduled, and the date of the last revision. Without one or more of these "time stamps," one cannot know whether content is current. Criteria for evaluating these dates and intervals between revisions vary greatly, depending on the nature of the content. Time-sensitive data require frequent verification and revision; other types of information are quite durable. Information that is not time-sensitive may be well served by static resources that are updated only infrequently or not at all. An important indicator of information integrity is knowing whether the provider or creator of the site is likely to maintain the resource over time. For example, resources developed by students, especially those developed as part of a short-term project, or by staff supported by one-time project funding, may not be maintained over time. The shelf life of such resources is frequently short, and the reliability of their content decreases at an accelerating rate over time. (See also Content Currency, Section 3.4.2.7.)

Site content that is revised periodically presents its own range of challenges. Whereas superseded information in print format remains accessible through the retention of earlier printings and editions, superseded Web content frequently ceases to exist. It is now, however, an increasingly accepted practice to archive superseded content when that content has potential durability or value over time.

3.4.4.2. Site Integrity. Site integrity pertains to the stability a site exhibits over time and to how a site is administered and maintained. It is, therefore, the responsibility of the site manager or Web master.

Even static resources require oversight, lest access to the resource itself ceases. Site integrity requires a commitment to ongoing maintenance. Broken links must be fixed promptly. Updates and revisions should be noted and reflected in revision numbers. Whether superseded content is archived, the URL where that content is maintained, and how requests to access the archive should be made is critical information that should be included in FAQs or through other forms of information pertaining to the site. Without adequate information on these issues, the integrity of a site will be poor or nonexistent.

3.4.4.3. System Integrity. System integrity is a measure of technical performance. The term refers to the stability and accessibility of the server hosting the resource. Assuring the quality of system integrity is the responsibility of the system administrator, not the creator of the information or the site manager. Server stability is perhaps the single most important criterion in judging the quality of a resource. The quality of the server ultimately determines the value of a Web resource. A server that experiences more traffic than it can handle will not provide consistent ease of access and will reduce the value of the resource, regardless of the value of its content. Frequent downtime or poor response time renders a resource virtually inaccessible. Anticipated downtime should be announced in advance, and all downtime should be reported, along with a projected date and time when the resource will again be available.

Important evaluators when judging system integrity are whether a site is mirrored and whether one can access it multiple times within a specified period (e.g., a certain number of times per day or week). Internet Scout Report staff, for example, check the availability of each site three times in the days prior to making it available. Staff responsible for selecting free Web resources at the Kuopio University Library in Finland require that links work sufficiently well on the third and final attempts to access a site over set intervals (Kuopio University Library Working Group 1996).

RECOMMENDED EXAMPLES OF SELECTION CRITERIA

Caywood, Carolyn. 1996. Selection Criteria for World Wide Web Resources. *Public Libraries* 60 (Summer):169.

DESIRE. Selection Criteria for Quality Controlled Information Gateways. Available at http://www.ukoln.ac.uk/metadata/desire/quality/toc.html.

HealthWeb. Selection Methodology and Guidelines. Available at http://healthweb.org/guidelines.cfm.

infor-quality-l. Selection Criteria for Internet Information Resources: A Poll of Members of infor-quality-l. Available at http://www.vuw.ac.nz/~agsmith/evaln/poll.htm.

Librarians' Index to the Internet. Selection Criteria for Adding

Resources to the LII. Available at http://www.lii.org/search/file/pubcriteria.

McGeachin, Robert. 1998. Selection Criteria for Web-Based Resources in a Science and Technology Library Collection. *Issues in Science and Technology Librarianship* 18 (Spring). Available at http://www.library.ucsb.edu/istl/98-spring/article2.html.

MedWeb. Guidelines for Inclusion of Sites in MedWeb. Available at http://www.medweb.emory.edu/MedWeb/history.htm.

Pratt, Gregory F. 1996. Guidelines for Internet Resources Selection. *College & Research Libraries News* 3 (March):134-135.

Scout Report. Selection Criteria. Available at http://scout.cs.wisc.edu.

WWW-vlib. Summary of Selection Criteria. Available at http://lists.w3.org/Archives/Public/www-vlib/msg00276.html.

4. ACCESS: RESOURCE DISCOVERY AND ADDED-VALUE FUNCTIONS

4.1. Resource Discovery

Many of the difficulties encountered by end users when searching the Web also confront library subject specialists and technology experts in their efforts to select free Web sites. Identifying high-quality Web resources is labor-intensive. Properly carried out, it is the most challenging and potentially most costly aspect of building scalable and sustainable collections. Although machine harvesting appears to be promising, it remains in nascent stages of development and is available only in limited settings. Resource discovery, evaluation, and indexing (i.e., cataloging) are still primarily manual processes that require well-formed strategies and efficiencies. The most clearly delineated resource discovery sources and strategies identified in preparing this report are those used by the Social Science Information Gateway. They include the following:

- joining discussion lists
- subscribing to distribution lists and e-mail publications
- monitoring and browsing sites
- *actively* searching the Internet
  - subject catalogs
  - higher education sources
  - Internet search tools
  - sites and lists that announce new Internet resources
  - Web agents
- searching non-Internet sources (e.g., scholarly journals, newsletters, Web reviews)

Resource discovery strategies and procedures outlined in the *DESIRE Information Gateways Handbook* (section 2-2) are also recommended.

4.2. Added Value: Cataloging, Metadata, Search Functions

Resource selection is not the only acquisitions function of libraries. To assure access, a library must provide the following range of value-added services:

- content description (e.g., descriptive and subject cataloging)
- resource organization (e.g., classification schema; indexing services)
- collection maintenance (e.g., provision of access over time, preservation, archiving, deselection)

These same responsibilities pertain to free Web resources. There is not yet full agreement on whether traditional cataloging practices (MARC, LCSH, MESH) and classification schema (Library of Congress or UDC) adequately describe digital formats or satisfactorily serve users. Some experts recommend the creation of MARC records stored in traditional OPACs, while others call for new methods of description and record storage. Regardless of the descriptive rules and type of catalog or database selected, there is a consensus that the minimum identification and retrieval data are as follows:

- title/name of resource
- location of resource (URL)
- author or editor (i.e., creator(s) of resource and of its intellectual content)
- publisher (i.e., organization making the resource accessible)
- free-text description, including audience

Other elements recommended for inclusion in the catalog record are those developed by the Dublin Core Metadata Initiative. They include the following:

- subject
- contributor
- date (created, last modified, data gathered)
- type (collection, database, guide/gateway, organization, service, home page, news service)
- format
- identifier
- source
- language
- relation (e.g., is part of, has part of, is a version of, replaces, is referenced by, is based on)
- coverage (geographic and temporal)
- rights

Unrelated to the specific retrieval data one should record and the format in which it should be recorded (e.g., MARC, Dublin Core), other cataloging issues need to be resolved. For example, if a resource points to other sites, one should determine whether each site requires its own unique record or whether a record for the "primary site" or collection is appropriate. Similarly, at what level is cataloging content adequate? What level of granularity should the cataloging record reflect? How these questions are answered will determine the quality of the collection. Other issues suggest that librarians may need to rethink traditional library cataloging practices, lest metadata cataloging backlogs equal or surpass the backlogs of uncataloged print resources stored in research libraries throughout the world.

RECOMMENDED EXAMPLES OF VALUE-ADDED SERVICES

Baruth, Barbara. 2000. Is Your Catalogue Big Enough to Handle the Web? *American Libraries* 31(August): 56-60.

DESIRE Information Gateways Handbook. 2.4 Cataloguing. Available at http://www.desire.org/handbook/2-4.html.

Dublin Core Metadata Initiative. Available at http://purl.org/dc/.

Humbul Humanities Hub. Describing and Cataloging Resources, version 1.0 (modified 20 Feb., 2001). Available at http://www.humbul.ac.uk/about/catalogue.html.

MacCall, Steven L., Ana D. Cleveland, and Ian E. Gibson. 1999. Outline and Preliminary Evaluation of the Classical Digital Library Model. In Knowledge, Creation, Organization, and Use. Proceedings of the 62nd ASIS Annual Meeting, Washington, D.C., October 31–November 4, 1999: Medford, N.J.: Information Today.

RENARDUS. Executive Summary. Available at http://www.renardus.org/deliverables/d6_1/doc0002.htm.

ROADS Cataloguing Guidelines. Available at http://www.ukoln.ac.uk/metadata/roads/cataloguing/.

Sowards, Steven W. 1998. A Typology for Ready Reference Web Sites in Libraries. *firstmonday Peer-Reviewed Journal of the Internet* 3(5). (See Elements in Typology of Ready Reference Web Site Designs, pp. 5-6.) Available at http://www.firstmonday.org/issues/issue3_5/sowards/index.html.

University of Virginia Libraries. 1998. Ad Hoc Committee on Digital Access. Final Report. Approved June 15, 1998.

5. DATA MANAGEMENT: COLLECTION MAINTENANCE, MANAGEMENT, AND PRESERVATION

Once created, collections of free Web resources require maintenance. Unlike print resources, free sites are highly dynamic. Content changes or is revised rapidly. In the context of higher education, superseded content can be crucial. In the print environment, superseded content can be easily retained or, if transferred to another location, relatively easily retrieved. The Web does not guarantee equivalent availability and access to superseded information. Unless archived and readily available through that archive, superseded information ceases to exist. Because of this ephemeral aspect of the Web, effective maintenance of collections of Web resources is labor-intensive and calls for a long-term staffing commitment. Collection maintenance can be among the more costly aspects in building and managing scalable and sustainable collections of Web resources. Archiving "snapshots" of Web resources as they exist at any given moment requires staff time and server space. Providing a mirror site for information developed and maintained by a third party is also costly, but not necessarily prohibitive, if the need to preserve or mirror is acknowledged to be within the scope of the collection. The National Library of Australia Pandora Project recognizes its obligation to provide indefinite access to the Web sites it selects for the project, and it archives these sites at the time they are cataloged. The largest and best-known archiving project is Brewster Kahle's Internet Archive, which employs Web-crawling robot software to collect Web pages from publicly accessible Web servers and examines links on these pages to locate, evaluate, and archive yet additional pages.

Maintenance tasks include the following:

- *Link checking*. Among the most persistent problems associated with collections of free Web resources are dead links. Various software programs are available to monitor links; these programs can be programmed to run at predetermined intervals. Recommended intervals range from once a week to once every three months. Longer intervals have proved counterproductive.
- *Reviewing error codes*. Perhaps the most frequently encountered error code is "403 Page not found." Resource locators frequently change, but the old URL may not point to the new location. Software will report these changes; staff members need to update all appropriate links.
- *Reviewing content*. Because content frequently changes, staff should regularly confirm that it remains consistent with descriptions found in cataloging records and that it continues to conform to the collection scope and policy.
- *Revising cataloging records*. Link checking and content review will determine whether and to what extent cataloging records should be revised.

- *Deselection*. Cataloging records and links should be deleted when a site can no longer be found or when its content no longer conforms to the collection scope and policy. Pointing to content that cannot be located discourages users from using the collection. Similarly, when the quality of a site has deteriorated or changed to the extent that it no longer meets users' needs or scope criteria, the site should be deselected by authorized staff.

RECOMMENDED EXAMPLES OF MAINTENANCE GUIDELINES

DESIRE Information Gateways Handbook. 2.6. Collection Management. Available at http://www.desire.org/handbook/2-6.html.

Humbul Humanities Hub.4. Collection Management. Available at http://www.humbul.ac.uk/about/colldev4.html.

National Library of Australia. Pandora Project. Available at http://pandora.nla.gov.au/.

Nicholson, Dennis, and Alan Dawson. BUBL Information Service: 8.5 Link Checking and Record Maintenance. In Wells, Amy Tracy, et al. 1999. *The Amazing Internet Challenge: How Leading Projects Use Library Skills to Organize the Web.* Chicago: American Library Association.

6. MULTILINGUALITY

The Internet knows no national or ideological boundaries. It permits users to access information, regardless of the country in which the server hosting the resource is located. Free Web resources are created and maintained in all the languages of the world. American research libraries collect foreign-language sites that support teaching and research in language and literature programs or certain subdisciplines in history, art, music, medieval studies, and political science. However, repeated informal surveys of free Web resources offered by leading American university libraries reveal that they neither reflect the breadth of non-English-language resources accessible via the Internet nor begin to approach the extent to which publications in languages other than English are represented in and continue to be acquired for their print collections. This is primarily because English is used so extensively in the Internet and because it is increasingly the language of choice for international communication. These two facts combine to encourage a regrettably large number of academics, including librarians and technical specialists, to underestimate the extent to which free Web resources with foreign-language content are needed to support higher education and research.

Previously, these sites presented major access problems for users, because software needed to display non-roman alphabets and character sets was not widely available. Inexpensive software programs

have now largely overcome this problem, but other challenges to incorporating non-English sites remain.

The British Resource Discovery Network (RDN) has addressed the issue of collecting non-English sites and recommends that inclusion should be based on appropriateness to the larger topic, scalability, and user demand. The RDN recommends

- a predefined number of languages that are significant or appropriate for the subject: importance of languages other than English for a specific subject (e.g., Danish for sites pertaining to Kierkegaard or Italian for sites pertaining to opera)
- value to user
- scalability: strategic language-by-language expansion of a site

The ease of displaying non-roman alphabets and character sets notwithstanding, including foreign-language content presents a range of additional challenges, including the following:

- Data presentation. What software and standards are required to display, search, and retrieve foreign languages using non-roman fonts? Should non-roman fonts be romanized?
- Metadata and cataloging rules. Should titles and the names of corporate bodies be translated into English? Should only English-language descriptive cataloging and keywords be used? Descriptions in two or more languages may enhance access but significantly increase workload.
- Searching and browsing. Should one expect to search by language or by domain name to narrow search results?

Many legitimately call for greater overarching foreign-language search capabilities. Research and development projects are currently under way to provide greater access to Web content without linguistic barriers through systems using cross-language information retrieval. The goal of these systems is to create search capabilities that permit the retrieval of sites to be independent of the natural language used to state the query. The success of such systems will depend on the broad application of emerging Web standards. The myriad issues and challenges pertaining to multilinguality lie outside the scope of this paper and will not be addressed here. Readers interested in these matters are referred to the following:

Oard, Doug W. 2000. Cross-Language Information Retrieval Resources (Overview) [last modified Nov. 24]. Available at http://www.ee.umd.edu/medlab/mlir/.

Peters, Carol, and Costantino Thanos. 2000. DELOS: A Network of Excellence for Digital Libraries; Promoting and Sustaining Digital Library Research and Applications in Europe. *Cultivate Interactive* 1 (July). Available at http://www.cultivate-int.org/issue1/delos/.

Koch, Traugott. 2000. Cross-Browsing in Renardus: Usage of Vocabularies in Renardus Gateways. Available at http://www.lub.lu.se/renardus/class.html.

RECOMMENDED EXAMPLES OF MULTILINGUALITY PRACTICES

Digital Asia Library (DAL). About DAL. Available at http://digitalasia.library.wisc.edu/about.html.

DESIRE Information Gateways Handbook. 2.12. Multilingual Issues. Available at http://www.desire.org/handbook/2-12.html.

Jennings, Simon. 2000. RDN Collections Development Framework, Version 1.1 (May). Available at http://www.rdn.ac.uk/publications/policy.html.

7. USER SUPPORT

Without publicity and promotion, a collection of Web sites can be an underutilized, even an unused, resource. A formal plan to inform potential users is essential. Publicity is best accomplished when collection creators identify their user groups and develop publicity and training materials best suited for those users.

Publicity can range from print media to electronic media and may include face-to-face presentations. In the digital environment, it may seem inappropriate to rely on print formats to promote Web-based resources, but flyers, posters, newsletters, articles and reviews in professional journals, and press releases remain the primary modes of advertising commodities and services. Print-based publicity is highly effective when directed to specific user communities, but traditional print formats have certain costs associated with production and distribution (e.g., paper and printing costs, distribution and advertising fees). Using e-mail for publicity purposes avoids these expenses, but staff must still be paid to prepare publicity. A good example of effective electronic publicity is the regular updates the Internet Scout Report sends to its list subscribers.

Face-to-face presentations, workshops, conference papers, and poster sessions can be highly successful, but the costs associated with such presentations (staff time, support to prepare presentation materials, conference registration fees, travel and lodging) mount rapidly.

RECOMMENDED EXAMPLES OF USER SUPPORT

DESIRE Information Gateways Handbook. Publicity and Promotion 2.8. Available at http://www.desire.org/handbook/2-8.html.

Internet Scout Report. Available at http://scout.cs.wisc.edu/scout/report.

SOSIG. Social Science Information Gateway. Training Materials and Support. Available at http://www.sosig.ac.uk/.

UKOLN. The UK Office for Library and Information Networking. Publicising Your Project. Available at http://ukoln.bath.ac.uk/services/elib/info-projects/publicise.html.

8. HUMAN RESOURCES: ORGANIZATIONAL AND FINANCIAL ISSUES

Collections of Web resources can be built by one person working in relative isolation or by large collaboratives working together on-site or at various locations. Staff may volunteer their expertise and services or they may be paid. There is no preferred model; each presents a range of options, advantages, and disadvantages depending on the scope and goals of the collection. No staffing model, however, can be successful if it does not recognize that building and maintaining the collection generate specific costs and present wide-ranging issues for communication and workflow across organizational units. In the case of libraries, a subject specialist working alone may have an impact on the workflow and priorities of the cataloging, reference/instructional, or technology staff. These costs and workflow issues must be recognized and addressed. The following section outlines staff skills and experience, training, individuals versus collaboratives, and costs associated with staffing and managing collections.

8.1. Staff Skills and Experience

8.1.1. Cataloging

Staff responsible for cataloging free Web sites benefit from broad training and experience in print-format description and subject catalogs. The Internet Scout Project in January 2001 posted a vacancy for a cataloger with the following skills and experience:

- Master of Library Science degree or corresponding experience
- educational/professional experience in electronic and networked information storage
- [Web-based] searching and retrieval
- knowledge of
  - AACR2
  - USMARC format
  - emerging standards, such as Dublin Core
  - Library of Congress subject headings

8.1.2. Selection

Skilled human involvement in the selection process is one of the most consistently called-for components of all projects studied in preparing this report. Harvester software may meet to a limited degree specific predetermined criteria for selecting resources, but only experienced subject experts (i.e., bibliographers, content experts, scholars) possess the level of knowledge required to select high-quality resources. Nonetheless, free third-party Web resources exhibit sufficiently different traits and characteristics from print and analog resources in terms of origin, content, authorship, access, and storage (archiving) that even the most skilled and experienced bibliographer of print publications would have difficulty in applying time-tested principles and guidelines for evaluating print and analog formats to scholarly resources in digital format. Other guidelines are needed. Some might argue that the selector trained in traditional collection-development practices is not the most appropriate person to identify, evaluate, and select free third-party Web resources. Some might even argue that the traditional bibliographer or selector is unprepared for the task at hand and that selection of Web resources more appropriately belongs in the realm of reference librarians, library technology staff, or other subject experts (e.g., advanced graduate students or faculty members). Such experience or training assures that those selecting for the collection understand user needs and expectations, and that they can base selection on a knowledge of the relevance and value of resources to the target audience. Subject experts are superior to harvest software because they can evaluate content critically and in a manner that harvesters have yet to master. Subject specialists should also be prepared to provide end-user training. Staff responsible for developing the intellectual scope and quality of collections should have experience developing analog collections or formal academic training in pertinent subject areas or both.

8.1.3. Technical Support

Central to successful Web resource collections is staff with excellent technical skills, regardless of the size of the collection. The role of staff is fundamental to the organization, access, and ongoing maintenance of the collection. Typical responsibilities of technical staff include the following:

- technical understanding of networked environment
- programming and scripting skills
- infrastructure software evaluation, selection, and maintenance
- interface development
- archival storage
- mirror site support (where appropriate)

8.1.4. Project Manager

The number and level of staff depend on the scope of the project. Large projects benefit from managers who can provide broad over-

sight and coordination. Persons with project management responsibilities should possess both subject and technical knowledge.

RECOMMENDED STAFFING SKILLS

DESIRE Information Gateways Handbook. 1.3. Staff and Skills Required Overview. Available at http://www.desire.org/handbook/1-3.html.

DESIRE Information Gateways Handbook. 2.1. Quality Selection. Available at http://www.desire.org/handbook/2-1.html.

Jennings, Simon. 2000. RDN Collections Development Framework. Version 1.1 (May). Available at http://www.rdn.ac.uk/publications/policy.html.

8.1.5. Advisory Boards

It is commonly accepted practice to appoint advisory boards to large projects and to those of extended duration. Board members should include subject specialists and technical experts. Their role is to shape the overall goals and objectives of the collection, to confirm that the project remains on course over time, and to address emerging issues.

RECOMMENDED ADVISORY BOARD MODELS

BIOME Special Advisory Group on Evaluation. Available at http://biome.ac.uk/sage/.

Edinburgh Engineering Virtual Library (EEVL). Annual Report to the eLib for the Period from 1st August 1995 to 31st July 1996. 1.2 Project Infrastructure. Available at http://www.eevl.ac.uk/document.html.

8.2. Staff Training

The nature of the Web and the characteristics of free Web resources challenge traditional collection-development and -management practices. This reality requires that staff receive training and supervision. The DESIRE Handbook recommends developing the following:

- exercises and examples for evaluating Web sites
- online tutorials
- staff manuals
- process to review sites selected by staff
- group e-mail lists to discuss and debate quality issues
- editorial meetings

Many quality sites are added to collections through user suggestions by means of "Contact us" or "Add new resource" buttons. Training users of such sites to become informed selectors is neither appropriate nor feasible; however, some means of quality control

should be maintained. The Humbul Humanities Hub has a policy that requires contributions from users whose credentials and selection criteria are unknown or cannot be judged to be reviewed, evaluated, and cataloged by staff.

RECOMMENDED STAFF TRAINING PRACTICES

DESIRE Information Gateways Handbook. 1.3. Staff and Skills Required Overview. Available at http://www.desire.org/handbook/1-3.html.

DESIRE Information Gateways Handbook. 2.1. Quality Selection: Training Staff. Available at http://www.desire.org/handbook/2-1.html.

Humbul Humanities Hub.3. Collection Management. Available at http://www.humbul.ac.uk/about/colldev3.html.

Jennings, Simon. 2000. RDN Collections Development Framework. Version 1.1 (May). Available at http://www.rdn.ac.uk/publications/policy.html.

8.3. Financial Issues

This report concerns building sustainable collections of free third-party Web resources—resources to which anyone can have full access without compensating the creator or host site. "End-user access without compensation" is the extent to which these resources are free. All value-added services that libraries provide to ensure improved access have significant costs. Value-added services for analog collections (e.g., selecting, describing, organizing, and storing) have specific costs associated with them. These costs are well-known to administrators and rather well documented in professional literature. Value-added services for free Web resources have similar costs; however, few people are aware of those costs, and comparative cost data, across collections or institutions, are not readily available.

8.3.1. Staffing

A good source for cost data is grant proposal budgets. One three-year project with a staff of 5.5 full-time equivalent (FTEs) estimated that total personnel costs would be $714,633 over the life of the project. The principal investigator's home institution agreed to cover the cost of equipment and software, which totaled $22,300. Excluding overhead, the total budget for this three-year project was $736,933. The principal investigator proposed developing a collection of "up to 10,000 sites," possibly fewer. The average cost per site in this project would be $73.69 if the project met its goal of 10,000 sites; the cost would be higher if fewer were selected. One might question whether the costs associated with this project reflect the average cost associated with building similar subject gateways. At the very least, this case demonstrates that there are identifiable costs associated with building collections of free Web sites. Except for the absence of a pur-

chase price, the nature of these costs closely resembles costs associated with acquiring, cataloging, and maintaining analog collections: staff costs for selection and description; technology costs for storage and retrieval.

8.3.2. Sustainability and Related Costs

Building sustainable programs of any kind requires moving from specific projects, which by their nature lie outside the scope of long-term institutional goals, to programs that are integral to the institution's goals and mission. How one elects to build sustainable collections of free, third-party Web resources has a direct effect on human resources, organizational models, and budgets. Staff members responsible for selecting and cataloging analog materials have full-time jobs. Increasing their responsibilities to include developing collections of free Web resources calls into question preexisting priorities (i.e., developing analog collections and other responsibilities). Creating opportunities for selectors to select free sites directly impinges on processing workflows. Which is a higher priority: processing new books that are not free, or cataloging free Web sites? If processing units receive additional staff to handle increased workloads stemming from the need to catalog free sites, have subject specialists received commensurate time to select and evaluate them? Further, what plans and provisions have been made to allow the requisite technical support of selectors' and catalogers' efforts? These questions underscore that new priorities in one area have a direct impact on workloads and priorities in other units.

Selecting free sites, whether a small number for inclusion in the OPAC or an entire collection to be maintained as a subject gateway, requires planning. Planning, in turn, requires that library managers understand and acknowledge that Web site selection is a new library-delivered service, or range of services, with specific and unique needs and with intrinsic and far-reaching implications. Workflows associated with analog collections run fairly independently of one another. After selection, the order is given to the acquisitions staff, who place the order, receive the item, and process the invoice before forwarding the item to the cataloging unit. Catalogers forward the item to staff, who apply call numbers, and other staff place the item on the shelf. There is rarely a need for cross-functional communication in the analog environment.

This is not the case with free Web sites. Each decision in the selection, cataloging, storage, and retrieval-interface process impinges on the process as a whole. Selecting free sites is a new responsibility. Thus, if existing staff begin selecting free sites, who will take on the work they previously handled? How will catalogers handle new formats for describing these resources? Can existing cataloging staff assume this responsibility without additional training, and who will continue the cataloging of analog materials? Can technical staff effectively provide access to these new resources? Do they have and understand how to use Unicode-compliant software? Can they cope with the new challenges of working with records based on Dublin

Core? These are substantive questions that libraries are struggling to answer. Understanding that there is no single right answer, and recognizing that every policy decision can have a direct and significant impact on work units that in the analog context coexisted without having an impact on each other's procedures or priorities will facilitate scalability.

Beginning with staffing choices and continuing through selection policy, cataloging practices, and interface design choices, building collections of free Web sets presents staff and administrators with a series of related issues that early in the planning process reveal themselves only superficially. For this reason, these issues require greater evaluation and consideration by all participants in the collection-building process and in looking at the process in its totality. Building collections cannot succeed if the process is viewed as a series of steps that coexist but do not influence or impede each other. Building collections of free Web resources must be viewed as a continuum—as a series of interdependent steps. Each component part has potential and probable influence and impact on one or more of the other parts. The scope of the collection influences selection, which, in turn, influences cataloging decisions. Technical limitations may determine the collection scope, cataloging practices, or other aspects of the collection. Understanding the range of issues and alternatives the collection will require and how they will affect each other will encourage the creation of multifunctional or cross-functional units that facilitate communication among those who must learn new skills (e.g., metadata formats) in order to provide new services (e.g., subject gateways). Undoubtedly, the staffing models created for project-based development of free Web sites will influence, if not determine, staffing needs and patterns for developing and maintaining sustainable collections of free Web resources.

8.3.3. Staffing Models: The Individual versus the Collaboratory

8.3.3.1. Individual Initiatives. Many outstanding collections have been built through the efforts of one person alone. Subject experts and collection curators are well positioned to identify resources relevant to their respective fields of expertise. The inefficiencies of this approach are numerous, but not necessarily so great as to rule out this approach in all contexts. Individuals can make important contributions if their collections are narrowly focused or specialized.

8.3.3.2. Departmental Initiatives. A library, a unit within a library, or a unit outside a library (e.g., an academic department) can quite effectively build collections. Subject pages and guides permeate home pages for libraries and academic departments. Even a cursory review of these sites will reveal a high degree of redundancy—75 percent or higher. This duplication of effort may not be "bad" or "wrong," but it should call for close consideration and evaluation. Departmental initiatives serve many purposes and may be highly successful within their specific context. Their greatest weakness may be that they reflect the traditional "pride of place" and institutional reputation that

have driven the building of print and analog collections—a reality created and encouraged by the nature of physical collections. Web-based resources do not have the fiscal barriers to access that characterize print-based materials. Why then build redundant collections that are unique only in their brand or URL?

8.3.3.3. Managed Collaboration. A review of stated and implied practices used in facilitating access to free Web sites suggests that the growth of these sites is too great to permit a single individual or institution to adequately identify and build collections in a timely manner. The Web is simply too vast. OCLC staff in 1999 described the number of Web sites as doubling annually; at the same time, half of all Web sites disappear each month. In other words, approximately 55 percent of all Web sites available on any given day did not exist one month earlier. Such statistics demonstrate the volatile and dynamic nature of the Web. If this growth and volatility continue, librarians will be well advised to emulate the collaborative Web harvesting projects of their colleagues throughout Western Europe and in Australia and New Zealand, where projects such as Resource Discovery Network (RDN), Social Science Information Gateway (SOSIG), Humbul Humanities Group, Finnish Virtual Library (FVL), EULER, and Pandora have advanced rapidly. Because these projects rely on collaboration among staff at multiple institutions and/or among special project staff, they have accomplished what no individual or single institution working in isolation can achieve: rapid and efficient collection development of nonredundant collections at a reasonable cost. In North America, the Internet Scout Report and the Digital Asia Library are two examples of specially funded projects staffed with full-time teams of subject specialists, technical experts, and metadata catalogers. These projects further illustrate that successful harvesting of high-quality Web sites is neither a part-time job nor an added responsibility for staff who are primarily accountable for other duties. In addition, discussions under way within ARL and various consortia underscore that successful mining of Internet resources will require libraries to provide users with vertical (i.e., deep) searching of Web content, not merely the horizontal (i.e., superficial) searching of sites typically provided by popular Web browsers (Campbell 2000).

8.3.3.4. Facilitated Collaboration. Facilitated collaboration is not based so much on shared principles, values, or aims as on the use of some high-level common framework for software such as DBOZ.org or the Cooperative Online Resource Catalog (CORC). The latter, organized by Netscape and others, is a collection of site reviews to which users may contribute. CORC is a metadata creation system for bibliographic records and pathfinders that describes electronic resources and has contributors from around the world. Both cases afford major benefits: large numbers of individuals coordinated by their home institutions contributing large numbers of sites, resulting in a rapid rate of collection development. Shortcomings include the lack of a single,

overarching set of selection criteria, limited assurance that resource descriptions reflect current content, and uneven subject coverage.

The highest level of collaboration is one in which participants recognize that decisions about metadata and controlled vocabularies need to be made, and that these decisions influence and determine collection scope, access, purpose (i.e., popular or scholarly), human resources, and cost. Who makes decisions and how decisions are implemented is fundamental to all forms and levels of collaboration.

9. FUTURE DIRECTIONS: NURTURING SUSTAINABILITY

Why collect free Web resources? The obvious answer is that current users need facilitated, value-added access to these resources to ensure that they will retrieve sites with high-quality content. The primary question for the future is whether broad application of enhanced metadata standards and next-generation search engines will allow end users to mine the Web themselves with greater precision than is currently possible and, in so doing, bypass the current need for facilitated access. In other words, will there be ongoing need for subject specialists (content experts) to provide the services traditionally provided by bibliographers and libraries?

For the foreseeable future, it is safe to say that the higher education community will remain dependent on collections of high-quality resources selected and described by experts using the practices outlined in this report. Near-term prognostications do not call for the subject expertise of humans to be replaced by computer-based search capabilities. Instead, the higher education community will grow increasingly dependent on free Web content made available through expanded human efforts to winnow, sift, and deliver access to a larger percentage of the Web's high-quality resources. Among the near-term future developments will be the following:

- increased outreach to user groups
- increased reliance on collaborative collection development
- greater emphasis on underrepresented subjects and non-text-based formats
- development of instructional support through course-specific collections or browsing by course number
- in-depth mining of distributed databases as foreseen in the Association of Research Libraries' scholars portal model
- simultaneous searching of analog and Web-based resources through the integration of distributed catalogs of Internet resources and library OPACs
- increased acceptance of internationally recognized cataloging standards
- increased control of URLs and descriptive metadata to reduce or eliminate broken links

- broad use of harvesting software to collect embedded metadata and thus facilitate the rapid cataloging of sites and eliminate redundant efforts
- decrease in manual harvesting and cataloging

Until now, building collections of free Web resources has been modeled on time-honored practices for building print and analog collections: an informed winnowing-and-sifting process that entails application of predetermined criteria and the exercise of human judgment. The Web is far too vast, its resources far too rich, for these same practices to prove successful over time. The size of the Web already exceeds human ability to review, organize, and manage collections at the level required to sustain the needs and priorities of higher education. The dynamic nature of the Web, particularly of its free resources, will render unviable manual review and cataloging. Research libraries are approaching an environment in which selection and cataloging of free Web resources will be machine-driven. Humans will develop selection criteria, but machines will apply them and accept or reject resources at a speed that only computers can deliver. Descriptive and subject analysis will be drawn from metadata embedded within the resources themselves by their creators.

Successful machine harvesting and cataloging techniques have yet to be perfected. The automatic methods that are currently under development, however, appear promising. The Library of Congress's Minerva project and the Swedish Royal Library Kulturarw3 project are examples of how sustainability of free Web resource collections will be achieved. How rapidly the process will be automated remains unclear. How easily automated procedures will be widely employed remains uncertain. Until technology can facilitate the harvesting and cataloging processes, manual practices will continue to be used and will be the foundation upon which a successful automated process is built.

Discussions of these topics and examples of current research projects in these areas are available at the following works and sites:

Arms, William Y. 2001. A Report to the Library of Congress: Web Preservation Project, Interim Report. Cornell University. Available at http://www.cs.cornell.edu./wya/LC-web/.

Campbell, Jerry. 2000. The Case for Creating a Scholars Portal to the Web: A White Paper. *ARL Newsletter 211*. Available at http://www.arl.org/newsltr/211/portal.html.

Dublin Core. Available at http://dublincore.org/.

Platform for Internet Content Selection (PICS). Available at http://www.w3c.org/PICS.

Resource Description Framework (RDF). Available at http://www.w3org/TR/REC-rdf-syntax.

Resource Organization And Discovery in Subject-based services (ROADS). ROADS Harvester software development. Available at http://www.ukoln.ac.uk/metadata/software-tools/.

Royal Library of Sweden. Kulturarw3 Heritage Project. Available at http://kulturarw3.kb.se/html/kulturarw3.eng.html.

10. REFERENCES

All Web addresses were functional as of June 5, 2001.

American Library Association. RUSA. BRASS Education Committee. 2000. Best of the Business Web Sites. Available at http://www.ala.org/rusa/brass/besthome.html.

American Library Association. MARS. Best of Free Reference Web Sites. 1999. Available at http://www.ala.org/rusa/mars/best1999.html.

American Library Association. ALSC. Children and Technology Committee. 1997. Selection Criteria: How to tell if you are Looking at a Great Web Site. Available at http://www.ala.org/parentspage/greatsites/criteria.html.

Arms, William Y. 2001. A Report to the Library of Congress: Web Preservation Project, Interim Report. Cornell University. Available at http://www.cs.cornell.edu./wya/LC-web/.

Arms, William Y. Collecting and Preserving Open-Access Materials on the Web: A Proposal to the Library of Congress from Cornell University. Unpublished report.

Baruth, Barbara. 2000. Is Your Catalogue Big Enough to Handle the Web? *American Libraries* 31(August):56-60.

Campbell, Jerry. 2000. The Case for Creating a Scholars Portal to the Web: A White Paper. *ARL Newsletter*, issue 211. Available at http://www.arl.org/newsltr/211/portal.html.

Caywood, Carolyn. 1996. Selection Criteria for World Wide Web Resources. *Public Libraries* 60 (Summer):169.

Chisholm, Wendy, Gregg Vanderheiden, and Ian Jacobs, eds. 1999. Web Content Accessibility Guidelines 1.0 (May 5). World Wide Web Consortium. Available at http://www.w3.org/TR/WCAG10/.

DESIRE. Selection Criteria for Quality Controlled Information Gateways. Available at http://www.ukoln.ac.uk/metadata/desire/quality/toc.html.

DESIRE Information Gateways Handbook. Available at http://www.desire.org/handbook/.

Digital Asia Library (DAL). Available at http://digitalasia.library.wisc.edu.

Dublin Core. Available at http://purl.oclc.org/dc/.

DutchESS. Dutch Electronic Subject Service. Available at http://www.kb.nl/dutchess.

Fedunok, Suzanne. 1996. Hammurabi and the Electronic Age: Documenting Electronic Collection Decisions. *RQ* 36(1):86-90.

HealthWeb. Selection Methodology and Guidelines. Available at http://healthweb.org/guidelines.cfm.

Humbul Humanities Hub. Describing and Cataloging Resources in Humbul. Available at http://www.humbul.ac.uk/about/catalogue.html.

IMesh: International Collaboration on the Internet Subject Gateways. Available at http://www.desire.org/html/subjectgateways/community/imesh/.

IMesh Toolkit. 2000. What is a Subject Gateway? Available at http://clark.cs.wisc.edu/imeshtk/.

Internet Scout Project. Available at http://www.scout.cs.wisc.edu.

Jennings, Simon. 2000. RDN Collections Development Framework. Version 1.1 (May). Available at http://www.rdn.ac.uk/publications/policy.html.

Kirriemuir, John. 1999. A Brief Survey of Quality Resource Discovery Systems. (Version 2, September). Available at http://www.rdn.ac.uk/publications/studies/survey/.

Koch, Traugott. 2000. Cross-Browsing in Renardus: Usage of Vocabularies in Renardus Gateways. Available at http://www.lub.lu.se/renardus/class.html.

Kuopio University Library Group. 1996. Selection Criteria for Virtual Libraries. Available at http://www.jyu.fi/library/virtuaalikirjasto/help/criteria.htm.

Lagoze, Carl, and David Fielding. 1998. Defining Collections in Distributed Digital Libraries. *D-Lib Magazine* (November). Available at http://www.dlib.org/dlib/november98/lagoze/11lagoze.html.

Librarians' Index to the Internet. Selection Criteria for Adding Resources to the LII. Available at http://www.lii.org/search/file/pubcriteria.

Library of Congress. BEOline: Selection Criteria for Resources to be Included in the BEOnline+ Project. Available at http://lcweb.loc.gov/rr/business/beonline/beonsel.html.

Library of Congress. Minerva: Mapping the Internet Electronic Resources Virtual Archive. Unpublished report.

MacCall, Steven L., Ana D. Cleveland, and Ian E. Gibson. 1999. Outline and Preliminary Evaluation of the Classical Digital Library Model. In Knowledge, Creation, Organization, and Use. Proceedings of the 62nd Annual Meeting of the American Society for Information Science, Washington, D.C., October 31–November 4, 1999: Medford, N.J.: Information Today.

McGeachin, Robert. 1998. Selection Criteria for Web-Based Resources in a Science and Technology Library Collection. *Issues in Science and Technology Librarianship* 18 (Spring). Available at http://www.library.ucsb.edu/istl/98-spring/article2.html.

MedWeb. Guidelines for Inclusion of Sites in MedWeb. Available at http://www.medweb.emory.edu/MedWeb/history.htm.

National Library of Australia. Pandora Project. Available at http://pandora.nla.gov.au/.

Oard, Doug W. 2000. Cross-Language Information Retrieval Resources (Overview) [last modified Nov. 24]. Available at http://www.ee.umd.edu/medlab/mlir/.

Peters, Carol, and Costantino Thanos. 2000. DELOS: A Network of Excellence for Digital Libraries; Promoting and Sustaining Digital Library Research and Applications in Europe. *Cultivate Interactive* 1 (July). Available at http://www.cultivate-int.org/issue1/delos/.

Platform for Internet Content Selection (PICS). Available at http://www.w3c.org/PICS.

Powell, Andy. 2000. RDN Terminology, Version 1.0. Available at http://www.rdn.ac.uk/publications/terminology/.

Pratt, Gregory F. 1996. Guidelines for Internet Resources Selection. *College & Research Libraries News* 3 (March):134-135.

Resource Description Framework (RDF). Available at http://www.w3org/TR/REC-rdf-syntax.

Resource Discovery Network. RDN Policy and Guidelines. Available at http://www.rdn.ac.uk/publications/policy.html.

Resource Organization And Discovery in Subject-based services (ROADS). ROADS Harvester software development. Available at http://www.ukoln.ac.uk/metadata/software-tools/.

Royal Library of Sweden. Kulturarw3 Heritage Project. Available at http://kulturarw3.kb.se/html/kulturarw3.eng.html.

Scout Report. Selection Criteria. Available at http://scout.cs.wisc.edu/.

SOSIG: Social Science Information Gateway. Available at http://sosig.ac.uk/.

Sowards, Steven W. A Typology for Ready Reference Web Sites in Libraries. *firstmonday Peer-Reviewed Journal of the Internet*. Available at http://www.firstmonday.org/issues/issue3_5/sowards/index.html.

Sweetland, James H. 2000. Reviewing the World Wide Web Theory versus Reality. *Library Trends* 48(4):748-768.

Thomas, Sarah. 2000. Abundance, Attention, and Access: of Portals and Catalogs. *ARL Newsletter* 212. Also available at http://www.arl.org/newsltr/212/portal.html.

UKOLN. The UK Office for Library and Information Networking. Available at http://ukoln.bath.ac.uk/.

University of Virginia Libraries. 1998. Ad Hoc Committee on Digital Access. Final Report. Approved June 15, 1998.

Waddell, Cynthia D. 1998. Applying the ADA to the Internet: A Web Accessibility Standard. Available at http://www.rit.edu/~easi/law/weblaw1.htm.

Wells, Amy Tracy, Susan Calcari, and Travis Koplow. 1999. *The Amazing Internet Challenge: How Leading Projects Use Library Skills to Organize the Web*. Chicago: American Library Association.

WWW-vlib. Summary of Selection Criteria. Available at http://lists.w3.org/Archives/Public/www-vlib/msg00276.html.

Additional Sources of Information

Agriculture Network Information Center. Available at http://www.agnic.org/.

Argus Clearinghouse. Mission. Available at http://www.clearinghouse.net/mission.html.

Cooperative Online Resource Catalog (CORC). Available at http://www.oclc.org/corc/.

DESIRE. Development of a European Service for Information on Research and Education. EU Project. Available at http://www.lub.lu.se/desire/desireIndex.html.

Engineering Electronic Library, Sweden. Available at http://eels.lub.lu.se.

European Libraries and Electronic Resources in Mathematical Sciences. The Euler Project. Available at http://rattler.cameron.edu/EMIS/projects/EULER/.

Internet Archive. Available at http://www.archive.org.

Renardus: The Clever Route to Information. Available at http://www.renardus.org.

WWW-VL History. Available at http://www. ukans.edu/history/VL/index.html.