Provenance, Journal of the Society of Georgia Archivists

Volume 20 | Number 1 Article 8

January 2002

What a Woven Web: Archives, Websites, and the Coming Legacy of "Light Gray Literature"

Earle E. Spamer
Academy of Natural Sciences of Philadelphia

What a Woven Web: Archives, Websites, and the Coming Legacy of "Light Gray Literature"

Earle E. Spamer

Website content is notoriously ephemeral. Its electronic existence is in communication with its components at one moment, gone at the next. A solution to preserving that content is to "permanently archive" the entire website. This raises concerns about technological accessibility and longevity. A website can also manifest itself as a dispersed collection of printed pages and downloaded electronic files redistributed amongst the paper and electronic records of individuals and organizations. What distinguishes that which is the record of an individual or an organization from the flotsam of reprinted and hyperlinked ephemera? Are archivists preparing appraisal methods for websites and their effluent?

First, can conventional appraisal methods be applied to website content? Can or should the electronic structure of a website be "arranged" if it does not have a hierarchical structure? Does the website structure lend itself to conventional descriptive methods? Can a website even be seen to be pre-arranged and self-describing? There is no consensus; the entire matter is largely overlooked.[1]

Second, archives and manuscript collections in their traditional functions will have to address the matter of appraising, arranging, and describing material that has been extracted partially, or copied completely, from websites. Such records may be working copies for reference purposes; others may be hardcopy printouts, "archival copies" of what has been posted electronically. Records may be received either digitally or as paper printouts; the format does not matter to the principle of provenance. The context of specific sets of records may provide the distinction. The demise of the original electronic host, however, turns an accumulation of downloaded and printed records into what could be the only documentation of the content of the website, even if it is a small portion of it. Such downloaded material would have been selected for a specific reason; it is not likely to be the same as a selection made by an archivist. To further complicate an archivist's view of such records, the authenticity of material out of the original website context is suspect if it does not contain an indication of original source. Archivists will have to be-

[1] The research presented here is based on the author's examination of fifty journals in archives management and library science, published between 1998 and 2001. In peer-reviewed literature, there are few substantive papers that address the matter of archiving entire websites, and none that consider archiving material extracted from websites. Websites are being archived now. To compare some conceptual differences between archiving websites and preserving electronic records, see the descriptions of the Clinton White House websites archived by the National Archives and Records Administration (http://www.clinton.archives.gov) and NARA's Center for Electronic Records (http://www.archives.gov/research_room/ center_for_electronic_records/about_the_center.html) that specifically archives "records designed *for* computer processing" (emphasis added here). Although the hyperlinks cited in this paper were current when the manuscript was written, by the time edited proofs were prepared six links had been modified and one website had disappeared. The links have been updated and are current as of July 2003. This example does not provide much optimism for documenting the authenticity of indexed links in archived website records. The archiving of websites has also been discussed at some length on the Archives and Archivists Listserv (to subscribe, send email to listserv@listserv.muohio.edu; in the body of the message write SUB ARCHIVES firstname lastname, or use the web interface at http://listserv.muohio.edu/archives/archives.html).

gin devising the protocols by which such records can be appraised.

WEBSITES AND ARCHIVES: SUMMATION OR SUBTRACTION?

Among an archive's many purposes is that it maintains the records of a single entity. In the current technological era, the contents of the official website of an organization might be properly construed as its own series of records. They result from the organization's activity of creating widely available information about it and produced by it. They are usually for public access, but parts of a website may be devoted to an organization's internal affairs. On the other hand, the website can be construed to be a single form of document, one that contains numerous sections and parts. Do records retention schedules apply to the website as a whole or to its parts? Once parts are deleted, unexpunged internal links to those parts are "dead."

Given the purpose and nature of the website, much of the information in it may be in abridged formats, selected from existing records, if not rewritten for brevity or clarity to a more public audience. Some documents might be an electronic text of a printed document or an exact facsimile of the document. A preserved "snapshot" of a website at a particular moment in time provides an arbitrary record of content and presentation, one in which selection has been determined by the creator, not the archivist.[2]

[2] NARA's "Clinton Presidential Materials Project" has already archived the first White House websites and made them available on its website (see note 1). Four versions (1994, 1995, 1996, and 2000) were created during the Clinton administration, each one of which was preserved. Embedded links in the websites are "dead," and not all images were provided to the National Archives. On the other hand, the National Archives' website includes functions to search all versions of the Clinton White House websites simultaneously, a kind of a finding aid unlike the serial approach of conventional finding aids. The original NARA press release (no. 01-34, 17 January 2001) is at http://www.archives.gov/media_desk/press_releases/nr01-34.html. (After this manuscript was written, NARA's website domain name was changed to www.archives.gov. Other page reassignments also made it difficult to relocate the original press release, which had been at page http://www.nara.gov/nara/pressrelease/nr01-34.html). Beyond the simple archiving of websites, already there is a hybrid website-archive that can be seen in DSpace at the Massachusetts Institute of Technology. It is meant to serve both as a traditional, widely accessible website, as well as its own archive of the work of MIT faculty and researchers (http://www.dspace.org).

From the perspective of an archive, the flow of information goes in two directions in the web environment. There is information placed there by an originating entity, and there is information extracted from it by other entities. The fact that information exists on a website makes it a candidate for transfer to an archive. How to accomplish this—technologically, with assurance of temporal longevity, utilizing conventional archival procedures of arrangement and description—is a current topic open to discussion and experimentation. So far, there has been no dialogue on this matter between documents creators and archivists.

The website has an unrestricted number of contact points and contained records. It is a paradoxical kind of document (or a series), one that is composed of ephemeral information. This concept is, ironically, well suited to the purpose of an archive. On the demise of a once-widely accessible website, its component records and their relationship to each other instantly become hard to identify and acquire. Unlike the "gray literature" of limited- and special-distribution documents and serials, so often difficult to locate even in their multiple copies, the electronic records of decimated and extinct websites are even more ephemeral—"light gray literature."

This is different from individual record loss or omission through selection in traditional archives and manuscript collections. Once disconnected from the web, the content of a website may still exist in one place (in an archive), but it is less likely to be accessed through the World Wide Web. Its pages will also be pocked by broken links and absent images. But even if the website were never archived and is utterly gone, it may yet exist as a constellation of randomly excised digital files and printouts lacking the perspective of original arrangement, fascicles of uncertain authenticity scattered through other archives and manuscript collections.

ARCHIVING WEBSITES

A set of authentic website records, electronically preserved as created and used by an organization, is the best means of verifying the form and function of the website.[3] But some administrators of an organization may tend to see the website as a means *to* archive information, with the fiscally driven good intention that if records are placed there, openly and widely available, it is cost-effective and inferentially long-lived.

Appraisal techniques might be confounded if a website contains borrowed material for which sources are neither documented nor credible. This may be a perceived problem only until standard methodologies are devised for appraising information contained in, and derived from, websites.[4] Sampling is not likely to be a satisfactory method of appraisal and accession. In fact, a selection of records that are a part of a website would seem to be contrary to the purpose of a website that is itself a selection of records.

A website may include evidential information, posted there as a means of making the information available widely and electronically. Such evidential data should already exist in other formal or legally sanctioned formats, as paper documents or as separately stored electronic files. The utility of having some evidential data available on a website may be a matter of convenience. So the primary purpose of archiving the content of a website is to preserve the informational aspects conveyed by its selective content and by the manner in which it was presented. This is an important criterion, one which archivists will have to take into consideration when appraising the content of a website, if the website is to be retained in "snapshot" format. One opin-

[3] Hardware and software obsolescence, and data migration to new media, are ignored here. These are important issues, but they are technology-dependent ones, the funding for which is an administrative issue. And if there is anything archivists have learned in the past forty years, looking back at how technology was seen and used can be supercilious. Forty years hence, current limitations should be no surprise.

[4] The matter has roots in cataloging and processing of electronic materials, in areas as diverse as classification terminology of web-based resources (e.g., Carol Jean Gody and Ray Reighart, "Terminology Identification in a Collection of Web Resources," *Journal of Internet Cataloging* 4 (2000): 49-65) to attempts to apply bibliographical description techniques to electronic resources (e.g., J. McRee Elrod, "Classification of Internet Resources: An AUTOCAT Discussion," *Cataloging and Classification Quarterly* 29 (2000): 19-38).

ion on the Archives Listserv summed up the predicament and process well: "View the web site as a business process, capture the information that most completely describes it[;] but who wants to keep an invoice forever[?]."[5]

There is a further element of subjectivity in a website, that of presentation. Depending upon the creativity and resources of a website's managers and operators, its content and presentation can range from a mundane, monochromatic posting of text documents, to a lively, colorful, complex series of interactive pages.[6] There may not be an understanding of how it was created or by whom.

The National Archives of Australia has established policies and procedures regarding the archiving of Commonwealth government public websites and its internal "webs" and "nets" of shared information and communicated documents.[7] Its procedures include not only the documentation of provenance and matters that fulfill legislative and fiduciary requirements, but they also provide direction for maintaining "records of web resource production and maintenance" and records retention appraisals. In addition, the directives specify that agencies must define and maintain a level of web-based recordkeeping that is adequate for its purposes.

The concept of *an* archive and archiv-*ing* is blurring. This is less of a conceptual misunderstanding than it is a reflection of how technological resources are used by people who are increasingly "information literate." Because information skills are increasing, the overall improvement of information literacy brings

[5] Archives Listserv, http://listserv.muohio.edu/archives/archives.html, Chris Flynn, 25 June 2002, responding to an inquiry from Marty Firestein, subject "Archiving websites."

[6] Michael J. Albers points out that "Web pages are no longer simple hand-crafted text objects, but dynamic groupings of text assembled moments before the reader views the page." Michael J. Albers, "The Technical Editor and Document Databases: What the Future May Hold," *Technical Communication Quarterly* 9 (2000): 191. While this may be true for many websites, many more are of the "mundane" sort. It is the more complex, "dynamic" kind that will be more susceptible to technological decay and inaccessibility in an archived state, at best reduced to raw text.

[7] National Archives of Australia, "Archiving Websites," http://www.naa.gov.au/recordkeeping/er/web_records/intro.html and http://www.naa.gov.au/recordkeeping/rkpubs/advices/advice43.html.

to the field of archival management a similar approval of, and confidence in, the application of technology in what not so long ago was very much a hand-driven, hand-arranged system.

This is all well and good for considering the future of work in archives. Yet although the broad value of archiving website-based information is recognized, there is not yet an equally broad response as to how to make it a part of archival methods. Some very large government organizations have accomplished it in some measure. By no means is website archiving moving ahead at the same pace as web resources are accumulating. The website is increasingly used as a means to make available, as much as possible, large subsets of an organization's records. Some kinds of records are unique to the website, susceptible to extinction without ever having had the chance for archival appraisal. This is a disheartening, disproportionate view of the potential for the website as a tool for archives management and as a source of archival information.

Website debris: archival or not?

Posting documents on a website makes them available. This is electronic document-management, outbound to users; it is different from managing documents coming to an archive. Posted documents do not ensure permanence, nor do they relate to an archive's purpose to make it possible for users to find sets of associated information. There also is no way of determining how many versions, revisions, abridgements, and copies exist. For a given document, too, users copy portions for their own use or redistribution. Some subset of such material may be records of the creator that are available nowhere else, but there is a great deal of website-based material that has been reposted from other sources. Source citations may be present in a document, but digital copies of documents with no embedded source line have little more acknowledgement than the date on which they were copied. On pages printed from the Web, many printing programs add a banner line citing the source's URL. On these banners a long URL is sometimes interrupted by an ellipsis, rendering the source citation useless. Certification that the material is unedited is almost never indicated. Copyright issues are dismissed as easily as they are acknowledged.

These are ethical issues, as well as procedural and legal ones. A large amount of material exists *as if* it belongs to the body of literature and resources in the public domain. Arranging records so derived from websites as a series of its own provides a logically descriptive set, but one that might be ignorant of the creators and arrangement.

NEW ARCHIVAL ISSUES RELATING TO WEBSITES

Electronic records have been concerns for archivists for some time. Now that archivists have to embrace website-based records—from downloaded individual web pages to entire websites—new concerns arise. Digitization of records has altered the ways in which basic research is done, now including the use of web-based resources. Primary sources are currently available widely, where once they were exclusively the domain of archives and (less frequently) published edited collections. Some organizations strive to meet this need by creating ways to make "unique" resources more available. For example, Rutgers University Libraries' Scholarly Communication Center is a web-based outlet created to "publish unique information sources on the Web that are not likely to be published elsewhere."[8]

There are shared-document "webs" and "nets" used internally by organizations. Their purposes are specific to the functions of the organization. Public websites, however, may contain many different kinds of records. Records with evidential and informational value are mixed. These records are widely accessible, copied and printed. Because of this, copyright and other intellectual-property issues are of significant concern to archivists.

The Society of American Archivists (SAA) first established its position on electronic documents in March 1995.[9] The position statement focuses on records that have been transmitted electronically; presumably this encompasses website-based data since in order to retrieve such records they must be transmitted

[8] Ronald C. Jantz, "Providing Access to Unique Information Resources: A Reusable Platform for Publishing Bibliographical Databases on the Web," *Library Hi-Tech* 18 (2000): 28. For a description of the Rutgers initiative, see Boyd Collins et al., *Building a Scholarly Communications Center: Modeling the Rutgers Experience* (Chicago: American Library Association, 1999).

[9] Society of American Archivists, "Archival Issues Raised by Information Stored in Electronic Form," *Archival Outlook* (May 1995), text also available at http://www.archivists.org/governance/handbook/app_j10.asp.

electronically. The SAA stated: "Electronic communications that are created, stored, or transmitted through electronic mail systems in the normal course of activities are records." They held the position that archivists should have authority to "determine the long-term value of [these] records" and that "significant changes in record keeping policies" are needed to retain and preserve these records.

In August 1997 the SAA provided a commentary on the management of intellectual property in the digital environment.[10] The comments responded to a National Humanities Alliance (NHA) statement, "Basic Principles for Managing Intellectual Property in the Digital Environment." The purpose of the SAA's commentary was to reinforce the positions of the NHA statement from the perspective of archives; it is not a critique or a reassessment of it. In February 1999 the SAA issued a statement on copyright issues relating to electronically distributed archival documents.[11] Together, these positions demonstrate that archivists are not "in the dark" about the important issues of ownership and authenticity of electronically derived records.

The authenticity of web-based documents that have been copied or downloaded from another source is compromised. Even web search-and-download processes have been dramatically automated. Commercial products are available for this purpose too. A standard methodology of research is the idea that working from original materials can withstand challenges raised regarding the materials' authenticity. The electronic environment lends itself all too easily to re-editing, substitution of materials out of context, and unacknowledged inclusion of other source materials.[12] The matter relates to textual and graphic materials alike, and further, to anything that is digitally recorded. The

[10] Society of American Archivists, "Basic Principles for Managing Intellectual Property in the Digital Environment: An Archival Perspective," http://www.archivists.org/governance/handbook/app_j4.asp.

[11] Society of American Archivists, "Statement on Copyright Issues for Archives in Distance Education," http://www.archivists.org/statements/distance_education.asp.

[12] See for example, Stephen Enniss, "The Role of the Artifact in a Facsimile Age," *RBM: A Journal of Rare Books, Manuscripts and Cultural Heritage* 1 (2000): 46-47.

opportunities for fabrication are limitless. For this reason, archives are best suited to serve as they always have, to arrange and preserve records that document its organization's activities. Because of an organization's internal control over its website, all such records carry a cachet of authenticity. The incorporation of copied or printed website documents into an archive is acceptable in the same manner as if they were manuscript materials, at the face value of those documents. Appraisal and arrangement procedures will apply. Impropriety will have to be addressed in the same fashion as would prevail if unauthentic or forged records were discovered in a paper-based archive. The medium should not be cause for administrative consternation.

As for the matter of archiving an entire website, it is the purest form of appraisal for an organization's archive to perform, even if the hypertext components are degraded by dead links and missing graphics. An organization's website is, to a point, pre-arranged and self-describing, a ready-made series of records (if not a single document containing many parts).

Copyright is a problematical consideration even in traditional environments of documents and records, particularly in manuscript collections. The problems were exacerbated when these concerns were applied to what was a non-traditional world of electronic records and communication. Now they are applied in a world that, for the most part, sees electronic records as equals of written records, but which is still grappling with the legalities of distribution. Numerous issues relating to web and other electronic copyrights are regularly discussed. A good summary by Charles Oppenheim[13] takes special note of the ease of copying and redistributing documents on the Web. He points out that there is a huge amount of unwanted linking to other web pages too that lays them open to copyists and downloaders, human and virtual alike. He suggests that copyright is "unlikely to survive in its present form."

FROM HERE, WHERE?

Everything that is usual in appraisal, accession, arrangement, description, and all matters of security and user services

[13] Charles Oppenheim, "Does Copyright Have Any Future on the Internet?" *Journal of Documentation* 57 (2000): 279-298.

is affected by the electronic environment. Archivists and data managers have been so dazzled by the processes, abilities, and economy of the Internet that they have rushed headlong to embrace it. As technology changes, the records are migrated to newer media and reformatted to remain readable–maybe. As Luciana Duranti opined on the understated but precise term "contemporary records," the missions and functions of archives, and the work and methods of archivists, will require some role changes.[14]

There is much to consider. A now-dated but usefully annotated "Bibliography on Electronic Records" lists many references that are applicable to management and duties of archives.[15] It supplements a 1993 annotated bibliography on the same subject by Richard J. Cox.[16] Together, these sources are a good historical introduction, showing the depth of work already done by 1996 in coming to terms with many problems of electronically created and distributed records. These works document a long period during which professional opinion, experimentation, and arbitration established the archivist's role in the management of electronic records. They model the process that can be followed to devise ways by which to professionally

[14] Luciana Duranti, "Meeting the Challenge of Contemporary Records: Does It Require a Role Change for the Archivist?" *American Archivist* 63 (2000): 7-14.

[15] Kimberly Barata, "Bibliography on Electronic Records," in *Functional Requirements for Evidence in Recordkeeping* (University of Pittsburgh, School of Information Sciences[SIS]). The lengthy bibliography is divided into thirteen sections, including theory, principles, and various issues of legal and professional practices. The bibliography had been posted on the SIS website at http://www.lis.pitt.edu/~nhprc/bibtc.html (last modified September 1996). In May 2002 the link was discovered to be a bad one, and the bibliography seemed to have been deleted from the website altogether. A recent search on the "Wayback Machine" of archived websites, accessible through www.archive.org, has relocated the missing bibliography at
http://web.archive.org/web/19991128184609/www.sis.pitt.edu/~nhprc/bibtc.html.

[16] Richard J. Cox, "Readings in Archives and Electronic Records: Annotated Bibliography and Analysis of the Literature," in *Electronic Records Management Program Strategies*, ed. M. Hedstrom (Pittsburgh: Archives and Museum Informatics, 1993): 99-156.

evaluate and manage websites and website-derived records transferred to the archive.

Archiving a website is really not much of a problem beyond the technical resources needed to retain it for continued use, even if it is no more than a collection of hyperlinked records. Keeping up with the technology is more of an issue for administrators; it is they who manage the funding and staff to continue certain practices. The challenge for the archivist is how to describe what is in that website. "Arrangement" is likely to be a fruitless (even futile) task. The nature of hyperlinks can preclude a sequential order to files. Series, as understood by archivists, may not exist in some websites. A website may also be a series unto itself, one with no orderly "folder list"!

Although files on a website will likely be listed by the creator in some kind of browsable fashion (but such a list may not have been created in the first place), access to the files can be gained from many *different* places on the website, as well as from anywhere on the Internet. Just how the website's creator organized and maintained the site will determine what, if any, kind of arrangement is discernible or not. Perhaps the simplest solution is to print out the site map as a finding aid of sorts. Of course, it will not show files that are nested within these first-level hyperlinks (akin to sub-series). If no site map exists, the archivist will have to be creative. Pragmatically there is not likely to be enough time to follow each link and all of its embedded links to exhaustion. And where does one stop? One step, or many steps, can end at a single document, and document-level control is not the principle objective of arrangement. Cross-linking to and from different places on the website makes it impossible to discern "subseries" within top-level links; some of them may be single documents too. It should be enough to summarize the website contents in a general description, leaving the navigation to the researcher or other archive patron. The best thing about this kind of environment is that the archivist need not fret about folders being inadvertently mixed, and concerns of theft or careless handling are practically moot points!

The issue of appraising and arranging downloaded and printed website records can be made easy or difficult. As with any photocopied materials that are contained in an individual's or organization's records, the same general principles can be applied. All such materials may be treated as if they were a collection of research materials to which a variety of copyright issues apply.

Unlike the photocopy, which inferentially means that the original or multiple identical copies may yet exist elsewhere, a downloaded web page may be unique if its website host has been switched off and not archived. The archivist may never know, nor should it be the archivist's job to find out. A new kind of reference service may come into being to meet the need to determine the "scarceness" of all of this light gray literature, a union list of sorts based upon URLs and document titles.[17] Surely these can be worked into encoded archival description (EAD) environments—accessible on a website, of course—and time will provide for the ultimate decision of whether or not such a service is practical and useful. Until a better realization is held of the volume of such material included in archived records, and until a better understanding is had of the intentions of records creators when they download web-based records, it may be better to err for a while on the side of conservatism, retaining more than what normally would be retained.

Perhaps just once in a generation archivists are in a position to establish standards by which a whole new technology and its records are treated. Now is that time.

Earle Spamer is archivist in the Ewell Sale Stewart Library of the Academy of Natural Sciences of Philadelphia, where he is also managing editor of scientific publications. He has worked in the curation of natural history research collections for thirty years, compiled the web-based "Bibliography of the Grand Canyon and the Lower Colorado River," published in several fields of history and science, and served as a contributing editor to the *Annals of Improbable Research*, where his work has also been translated into German, Italian, and Chinese.

[17] A "Bibliographic Object Name Resolver Service" is used within the University of Michigan's Humanities Text Initiative website, http://www.hti.umich.edu. For SGML documents posted there, bibliographical information is provided, complete with its URL fully spelled out, with verification that it is a persistent URL. Although this is for printed works that have been digitally scanned and made available through the Web, a comparable listing might be desirable for URLs themselves.