Committing the Web to Memory: Transmitting Web-based Records over Time

Jim Suderman

This paper will draw upon the findings of an InterPARES 2 case study on website exhibits. , , are at risk due to The interaction between the creator's platform and the rendering of the record by the user's platform results in a risk to the meaning and intent of web-based records, perhaps even more than to other electronic records The paper will examine processes that result in the creation and maintenance of web-based records in terms of record values appropriate to their context. These values define the characteristics of record authenticity, reliability and accuracy. From this base the paper will conclude with an exploration of the usefulness of web-based records as evidence over time.

The title of this paper, "Committing the web to memory," plays on a phrase that represents what for me was a school activity – memorizing information, like Hamlet's soliloquy, to produce later in a performance or a test – and something we probably do every day: save a document to computer memory to access it later. The inventor of the world wide web, Tim Berners-Lee, defines the web as "The set of all information accessible using computers and networking, each unit of information identified by a U[niversal] R[esource] I[ndicator]."[1] This definition focuses almost exclusively on the technology: computers, networks, universal resource indicators. The exception is the reference to units of information. Identifying Hamlet's soliloquy by a URI may be effective for moving it around networked computers, but all other aspects of its cultural context must be supplied by the viewer. It is safe to say that Shakespeare did not compose the speech only for it to be accessed by networked computers. For this reason I suggest expanding the preceding definition of the web to qualify "units of information" by the phrase "developed for a specific purpose" and made accessible using computers and networking.

The principles that guide archival practice are designed to preserve the intent of the creator over time for the benefit of subsequent users of the creator's records. The code of ethics

[1] From the glossary of Tim Berners-Lee with Mark Fischetti. *Weaving the Web. The Original Design and Ultimate Destiny of the World Wide Web by its Inventor.* (San Francisco: Harper, 2000). Definition available at http://www.w3.org/People/Berners-Lee/Weaving/glossary.html (checked 14 March 2005).

adopted by the International Council of Archives requires archivists to preserve and provide access to records in such a manner that protects "the integrity of archival material and thus guarantee that it continues to be reliable evidence of the past." [2] This statement is consistent with a 1991 resolution of the European Council recognizing archives as having a twofold purpose, as "a basis for decision-making in the public sector on the one hand, and as a vital component of a nation's cultural heritage on the other."[3] It is valid, then, to examine how archival principles operate in relation to the preservation of web-based records – governmental, scientific, or artistic.

The word "archives" in the European resolution is a reference to records: units of information with a specific purpose. To be properly understood records must be considered in the context in which they were created. For decision-makers to be accountable the records they use must be accurate and reliable. In terms of cultural heritage, de-contextualized information cannot act as a vital component because at the very least the cultural context must be established for records to act as a vital component of a nation's cultural heritage. For these reasons the two purposes in the European resolution are useful starting points when approaching web-based records for the purpose of preserving and reproducing them in the future. For web-based information transmitted across space and through time to stand as evidence of the intentions and purposes of its creators requires preservation activities that identify the web-based content with its creators and their purposes and intentions.

[2] International Council of Archives. *Code of Ethics* (1996). Available at http://www.ica.org/biblio/code_ethics_eng.html. (Checked 14 March 2005.)
[3] "Resolution of the Council and the Ministers of Culture, meeting within the Council of 14 November 1991 on arrangements concerning archives." *Official Journal C 314, 05/12/1991 p. 0002-0002.* Available at http://europa.eu.int/smartapi/cgi/sga_doc?smartapi!celexapi!prod!CELEXnumdoc&lg=EN&numdoc=41991X1205(01)&model=guichett. (Checked 14 March 2005.)

This paper examines some of the means by which the creator's intent and purpose can be ascertained and represented and how some current initiatives, in particular the Internet Archive, approach the preservation of web-based information. The first part draws heavily on the findings of a case study on the creation, maintenance and preservation of web exhibits conducted within the InterPARES 2 Project. This research project is examining authenticity, reliability and accuracy of records created within emerging digital systems.[4] An envisioned outcome of the Project is a principle-based policy framework that will provide preservation guidance for creators and memory institutions responsible for preserving the accessibility of accurate, reliable and authentic records over time.

The case study focussed on web exhibits created by two Canadian archival institutions. Individuals fulfilling the roles of manager, website coordinator, scanning technician and exhibit curator were interviewed to highlight the purpose and process of creation, the technology involved, and the on-going record keeping requirements.

Both subject institutions are required to make their archival holdings known to as wide an audience as possible. Web exhibits are seen as an important and cost effective means to fulfil this responsibility and both institutions devote significant resources to the development of web content in general and web exhibits in particular. Although web exhibits are not transactional

[4] This research project is examining authenticity, reliability and accuracy of records created within emerging digital systems and is directed by Dr. Luciana Duranti, University of British Columbia. See www.interpares.org for general information on the InterPARES 2 Project and its predecessor project. The InterPARES 2 Project proposal defines accuracy as "the truthfulness of the content of the record and can only be established through content analysis. With administrative and legal records, it is usually inferred on the basis of the degree of the records' reliability and is only verified when such degree is very low." Authenticity is defined as "the trustworthiness of a record as a record. An authentic record is one that is what it purports to be and has not been tampered with or otherwise corrupted. Authenticity is established by assessing the identity and the integrity of the record. It must be possible to ascertain at all times what a record is, when it was created, by whom, what action or matter it participated in, and what its juridical/administrative, cultural, and documentary contexts were. It must also be possible to ascertain the wholeness and soundness of the record: whether it is intact or, if not, what is missing." Reliability is defined as "the trustworthiness of a record as a statement of fact. It exists when a record can stand for the fact it is about, and is established by examining the completeness of the record's form and the amount of control exercised on the process

records in the same way as contracts or registrations, the study showed that considerable efforts are made to ensure that exhibits are accurate, authentic and reliable. The study identified three more or less concurrent processes in the creation of web exhibits: research, administrative, and technological. Each contributes to some aspect of ensuring that the exhibits were trustworthy records of the institution.

Research process

The research process for creating web exhibits is based primarily on scholarly research practices. Once an exhibit topic or subject is determined and approved, the curator reviews existing secondary literature. This review provides a basis for a preliminary narrative for the exhibit topic and assists in identifying items for inclusion in the exhibit. Items for inclusion are predominantly digital images of conventional, that is paper-based, records from the holdings of the subject institutions. Items are chosen on the basis of

- their authenticity
- their ability to represent a variety of narrative viewpoints
- their value for quotations that contribute immediacy and provide context to the narrative
- and visual appeal (e.g., maps or technical drawings).

The initial identification of items for use as sources includes compiling citations to assist with their review and final selection. At the same time as sources are being selected the narrative is refined and images of source documents, quotations and citations begin to be integrated.

Administrative process

The study identified two key managerial approvals. The first was the initial approval for the exhibit concept or topic. This approval clearly establishes institutional ownership of the

---

of its creation." See Luciana Duranti, *InterPARES 2 Project Detailed Proposal*, p. 1.1-11, available at http://www.interpares.org/display_file.cfm?doc=ip2_detailed_proposal.pdf. (Checked 14 March 2005).

exhibit process and allows the research and technological processes to begin. The final approval, by the institutional head, marks the end of the process of creation; it thus marks the completion of the record, the institution's satisfaction with it and the acceptance of the institution's ownership and responsibility for it. Because exhibits often portray historical events, one consideration in the final approval is to ensure that events are not portrayed in a partisan or otherwise inappropriate fashion.

There are also approval steps during the process of creation. These govern factors such as the focus of the narrative, the selection of images and text, the allocation of resources – for example, approval is needed if identified source materials require extensive conservation work prior to scanning – and the 'look and feel' of the technological components of the exhibit. Interim approvals may also involve a review by other staff to help ensure topical accuracy or clarity of presentation.

Technological process

The technological process involves two sub-processes. The first is the creation of the digital components used within the web exhibit and is a process that is shared by the website coordinator and the scanning technician. The second sub-process is the creation of the web exhibit itself, normally the responsibility of the website coordinator.

Creating the digital components predominantly means scanning or digitising source materials into image files, but may also include the digitisation of video files or recording of sound files. The scanning process was developed independently of web exhibit creation as an accessibility and preservation initiative for photographic holdings, including negatives and rare formats, in both institutions. The scanning technician creates high quality tiff format[5] images

[5] Tiff stands for Tagged Image File Format, a popular, flexible and public domain raster file format. The tiff specification is available at http://partners.adobe.com/public/developer/tiff/index.html. (Checked 9 March 2005).

from which images in jpeg format[6] are derived for use as web exhibit components. The technician's visual skills are supported by high quality scanning software and hardware, including calibrated monitors for correct display. In the creation of the digital components for web exhibits the same high degree of accuracy is sought as for the reproduction of photographic records for access and preservation purposes. The relationship of the scanned image to the source item and any other records it may be related to is maintained by the contextual data added to each image. Details of the scanning process, including technical settings, are likewise maintained as attached data elements. The contribution of the rigour of this process to exhibit components is analagous to that which the scholarly research process brings to the development of the exhibit narrative.

The second technological sub-process is the assembly of the digital components comprising the exhibit content into the actual exhibit. Exhibits consist of a number of linked, HTML-encoded pages; each page is comprised of a number of digital files including the HTML code file and image (and other) files needed to complete its content. A style sheet governing type style, font, color and other visual and structural aspects is used for all web pages comprising the website.[7] Web pages are constructed according to a standard corporate web page template that contributes considerably to the identity of the record through logos, navigational links and copyright statements. Another corporate standard requires that all web content, including exhibits, must meet accessibility guidelines specified by the World Wide Web Consortium.[8]

---

[6] The jpeg [Joint Photographic Experts Group] format is a compression format for images and is an ISO standard *Digital Compression and Coding of Continuous-tone Still Images, Part 1: Requirements and Guidelines*. (ISO/IEC IS 10918-1). The jpeg format appears to have a baseline specification, from which additional extensions are added, which may or may not be supported by supporting applications. Specification information is available at "JPEG File Interchange Format" at http://netghost.narod.ru/gff/graphics/summary/jfif.htm. (Checked 9 March 2005.)

[7] Definition of cascading style sheets and specifications are available from the World Wide Web Consortium, at http://www.w3.org/Style/CSS/. (Checked 9 March 2005.)

[8] World Wide Web Consortium, "Web Content Accessibility Guidelines 1.0", available at http://www.w3.org/TR/WCAG10/ (Checked 9 March 2005.)

These requirements include accurate text representations of images so that browser software can "read" the text describing an image to blind visitors to the exhibit, for example. This process not only requires the correct assembly of all the exhibit components but also checking to ensure that they will display correctly on the most commonly used software and hardware platforms.[9]

The exhibit is finalized on a development server that emulates the actual environment encountered by an exhibit visitor. While residing on the development server exhibits are checked using the designated browsers to ensure proper display. Once the exhibit receives final approval the website coordinator forwards it to the corporate service provider where it is up-loaded to the production server. It is from this server that visitors, including staff of the creating institutions, actually access the exhibits. For reasons of security and efficiency, the production server is maintained centrally for all Government offices and is therefore not in the control of the institutions developing the exhibits.

Each of the three processes[10] contributes to the accuracy, authenticity, and reliability of the web exhibits. In terms of accuracy, scholarly research practices contribute directly to the "truthfulness" of exhibit content through the critical use of sources and source citations to develop the exhibit narrative, i.e., the critical use of sources provides independently verifiable facts on which the narrative is based. To gain managerial confirmation that accessibility requirements have been met, textual representation of non-textual content must be accurate. Skilled staff, high quality tools, and an established procedure for digitizing images contribute to the accuracy of the image components of web exhibits. Testing how the web exhibits display on several common platforms helps ensure that the exhibit will be accurately rendered for visitors.

[9] Corporate web log statistics are used by the two institutions in the study to determine the most commonly used software and hardware.

The curatorial process enhances authenticity through the selection of authentic sources to include in the exhibit. This means sources whose identity and integrity can be shown to be intact. Executive level approval of the finished exhibit provides a clear indication that the institution accepts the record as being what it purports to be.[11] The procedures and tools for scanning images transfers the authenticity of the selected source records to the digitized copies. Usage of the corporate web template on each page is significant because the template elements identify the exhibit as belonging to the creating institution.

Control over the entire process of exhibit creation is evident from the initial and final managerial approvals. The final approval is also an indicator that the record is complete. On-going managerial review during the exhibit process serves as a check that scholarly research and other required practices are being followed. The established procedure for scanning results in the production of digitized exhibit components that are complete and reliable. Finally, there is the testing of the assembly of digital components into web pages, and pages into exhibits by the website coordinator within a development or test environment. These elements of the three processes contribute to the reliability of the web exhibits as records.

The study also revealed noteworthy gaps in accuracy, authenticity and reliability both in the rendering of the record, and in relation to institutional record-keeping and security. Web exhibits are not integrated into any existing record keeping systems. While the website itself could be developed as a record keeping system, it currently lacks key characteristics that would make this possible. Notably absent are a file classification plan or other tool that establishes the relationship between web exhibits and the other records of the institution, and defined record

[10] The three processes outlined here are consistent with those defined by Martin R. Kalfatovic, *Creating a Winning Online Exhibition. A Guide for Libraries, Archives, and Museums* (Chicago and London: American Library Association, 2002), p. 20.

retention requirements that set out how long the exhibits must be maintained. If the web exhibits were to be removed from the context provided by the institutional web site, their trustworthiness as institutional records would suffer. The degree of loss of trustworthiness might be mitigated at least partially by compensatory mechanisms such as a preservation plan for the records.

Web standards and security are outside the ambit of the subject institutions in the study. Web standards, such as the requirement to use the web page template, and security, including the maintenance of the live or production web servers, are both centralized within the corporation. There was no evidence to suggest that security procedures and policies had been developed in relation to specific institutional activities or record requirements. The institutions under study followed no special procedures to ensure that web exhibits were not tampered with or modified after being up-loaded to the production server.

Since exhibits are developed specifically for users or clients external to the institutions, it is worthwhile to note deficiencies in terms of rendering the exhibits. Corporate standards require only that web content display properly on the most common browsers with minimal display requirements: that images display with a 256 color palette, for example.[12] Emerging technologies such as cell phones can access web content with non-standard browsers and have very limited display capabilities. A cell phone tested within the study failed to render several components of web exhibits. Generally speaking, it successfully displayed text and the required navigation links. However, neither the graphic components nor their alternative text description were rendered. As a result, all illustrations, including the institution's name and other identity

[11] This reflects authenticity as it is defined within the ISO 15489 standard: *Information and documentation – Records management – Part 1: General*, (2001), p. 7, section 7.2.2 "Authenticity."
[12] Government of Ontario. *GO-ITS 23.1 – Internet Public Access – Product Design*. Sections 1.7.8 and 1.7.9 (2002), available at restricted intranet site http://www.gov.on.ca/MBS/techstan/GOITS_23_1_Internet_Public_Access_Product_Design.htm. (Checked 14 March 2005.)

components, were not displayed because this information was all formatted as graphics. When conventional monitors and browsers were used to access the exhibits off-line, one browser failed to locate or apply the style sheet. The result was that some components, such as the navigation links, appeared but did not display correctly. For the viewer, improper rendering of exhibits will negatively affect the accuracy, authenticity and reliability of records. As technology continues to evolve rapidly this may negatively affect the accuracy, authenticity and reliability of the exhibits *even for the creator* unless information such as an optimum viewing platform is identified or some other means of illustrating what should appear is provided.

It is important to analyze what makes web-based records accurate, authentic, and reliable especially as web-archiving initiatives are already underway. Brewster Kahle's founding vision for the Internet Archive is to provide universal access to all human knowledge. Accessibility is, indeed, a key component in the mandates of most "memory" institutions like archives and libraries. And while it is worth noting that nowhere in the Internet Archive's *Terms of Use* do the terms "preserve" or "preservation" appear, preservation is, nevertheless, implicit. For example, the proviso that "according to standard academic practice, if you use the Archive's Collections for any research that results in an article, a book, or other publication, you list the Archive as a resource in your bibliography" implies that the cited source will be preserved so that others may check it for themselves. The *Terms of Use* go on to say that "the Archive makes no warranty or representation regarding the accuracy, currency, completeness, reliability, or usefulness of the content in the Collections" nor "that access to the Collections will be uninterrupted, timely, secure, or error free, or that defects, if any, will be corrected."[13] These

[13] Internet Archive, *Terms of Use* (10 March 2001), available at http://www.archive.org/about/terms.php, accessed 20 October 2004.

statements absolve the preserver of any responsibility to creators and users alike for maintaining and providing access to trustworthy records.

My purpose here is not to belittle or defame Mr. Kahle's laudable initiative but rather to contrast the idea of records as units of information created for a specific purpose with those which are simply identified by a URI and accessed using networked computers. The Berners-Lee definition of the web is very much reflected in the Internet Archive, which is designed first and foremost to make accessible whatever it ingests. Information that is not restricted is harvested and maintained only in the context of date and URI. The Internet Archive's acquisition process does not include collecting information on website creators identities or intentions and so if this information has not been made explicit in the web content they will not be evident in the Internet Archive.

Contrast the Internet Archive with author- or creator-focussed preservation initiatives that have existed for some time, especially in the library field where national libraries have sought to extend depository programs to include digital, including web, publications. Library and Archives Canada asks depositors to send them an email if they have something to deposit and promises to maintain the "integrity and authenticity of their online publications for future generations."[14] A more ambitious project is the PANDORA initiative of the National Library of Australia which takes responsibility for capturing, archiving and providing long-term access to selected online electronic Australian publications.[15] The purpose of this brief comparison of two approaches to archiving web-based records is to highlight how the fundamental differences in the two general approaches affect how what is gathered is maintained for preservation.

[14] Library and Archives Canada website, *Electronic Collection*, "Depositing to the Electronic Collection", available at http://www.collectionscanada.ca/electroniccollection/003008-200-e.html. (Checked 14 March 2005.)
[15] See "Pandora Project" at the National Library of Australia website, http://www.nla.gov.au/policy/plan/pandora.html, accessed 10 March 2005.

The web is, by definition, a means of transmitting of information across space. Committing the web "to memory" is a commitment to transmitting information through time as well. For web content to move from being simple units of information to records that can support accountable decision-making or contribute to a nation's cultural heritage it is critical that such information be identified in terms of its context of creation. The role of creators in the preservation of meaningful web content is as important as that of preservers. However, the case study suggests that while considerable effort is made in the creation of web-based records, much of the value achieved thereby is attenuated by the absence of record keeping and preservation processes to maintain the records.

Preservation activities must include providing a description the creator's purpose and intent both in the process of creating web-based records. The preserver must also document how records are maintained and reconstituted to counteract erosion of the accuracy, authenticity and reliability of those records over time.

I began this paper with a play on the words "committing to memory." A technology based preservation initiative, such as the Internet Archive, would treat Hamlet's soliloquy as an "information unit" beginning with "To be or not to be" and ending with "Be all my sins remember'd." Its identification would simply be a URI, e.g., http://www.wowzone.com/hamlet.htm. This is sufficient for the purpose of moving it across space using networked computers but is insufficient for accountable decision-making or contributing to a nation's cultural heritage. An author or creator-oriented preservation of the same text would preserve its relationship to a character named Hamlet in an eponymous play by a sixteenth century English playwright named William Shakespeare. This approach supports decision-making, by establishing that the soliloquy is part of a larger work, for example. This

approach also makes clear to whose cultural heritage the speech contributes and is key therefore in moving information through time.