InterPARES 3 Project

International Research on Permanent Authentic Records in Electronic Systems

Title: Case Study 09 – Alma Mater Society of the University of British Columbia: Policies and Procedures for Web Site Preservation

Case Study Report

Status: Final (public)

Version: 1.3

Date Submitted: November 2009

Last Revised: May 2013

Author: The InterPARES 3 Project, TEAM Canada

Writer(s): Helen Callow
School of Library, Archival and Information Studies,
The University of British Columbia

Brian Sloan
School of Library, Archival and Information Studies,
The University of British Columbia

Elizabeth Shaffer
School of Library, Archival and Information Studies,
The University of British Columbia

Project Component: Research

URL: http://www.interpares.org/ip3/display_file.cfm?doc=ip3_canada_cs09_final_report.pdf

Document Control

Version history			
Version	Date	By	Version notes
1.0	2009-11-09	H. Callow, B. Sloan, E. Shaffer	Discussion draft prepared for TEAM Canada Plenary Workshop 05.
1.1	2009-11-23	E. Shaffer	Incorporation of feedback received from S. Goldfarb.
1.2	2009-11-24	R. Preston	Minor content and copy edits.
1.3	2013-05-04	R. Preston	Minor content and copy edits for public version.

Table of Contents

Case Study Report

A. Overview

The Alma Mater Society (AMS), located on the Point Grey campus of The University of British Columbia (UBC), is the University's student society. Founded in 1915, the society consists of close to 44,000 members made up from students at the Vancouver campus and students at UBC's affiliated colleges. In 1928, students incorporated the AMS as an independent non-profit society under the British Columbia *Society Act*.

The AMS oversees services to students (tutoring, job hunting, etc.), businesses and clubs. The AMS Archives is the archives and records centre for the Alma Mater Society.

In November 2006, the Society's Archivist, Sheldon Goldfarb, approached the InterPARES 3 Project to join as a test-bed partner and proposed a records case study in a document dated November 2006.[1] The study examines the Society's Web site with a view to determine strategies for the long-time preservation of a Web site that is frequently changing. The archivist was interested in developing strategies for exercising greater control over modifications to the Society's Web site, and for the long-term preservation of its various iterations over time.

This final case study report is presented to TEAM Canada, and incorporates the final decisions made by the AMS and an Action Plan that devises strategies for control and long-term preservation of its Web site.

B. Statement of Methodology

The methodology used in conducting research for the AMS case study is known as Action Research. Action research is a collection of participative and iterative methods, which pursue action (in this case, the preservation of a digital Web site) and research at the same time. As a matter of course, action research forges collaborations between community members and researchers in a program of action and reflection toward positive change.[2] Action research makes extensive use of case study methodology and of direct communication and interaction with

[1] See www.interpares.org/ip3/display_file.cfm?doc=ip3_canada_ubc_ams_research_proposal.pdf.
[2] Greenwood, David J. and Morten Levin, "Reconstructing the Relationships between Universities and Society through Action Research," in Norman K. Denzin and Yvonna S. Lincoln, eds. *The Landscape of Qualitative Research: Theories and Issues*, 2nd (Thousand Oaks: SAGE Publications, 2003), 131-166.

subjects of the research, who are at the same time participants and contributors in the research activity.

The AMS's Web site was identified as a body of digital material for which a preservation plan will be developed. Data were collected about the institution's context and limitations, the specific body of material, its documentary forms, technological constrains, and the functional and cultural meaning of the materials.

The Graduate Research Assistants worked closely with the AMS Archivist to complete the study. As required by the procedures of InterPARES 3, information regarding the institution, its records and its operations was compiled through an ethnographic approach to the study. Various interviews and observations were conducted with the Society's Archivist, Communications Manager, Web site Editor and its Information Technology Manager, producing the contextual analysis, diplomatic analysis and providing responses to the records case study research questions, and to gain a cultural perspective of those responsible for the Web site.

As a result of the submission of these three documents to the researchers at the May 2008 TEAM Canada Plenary Workshop, the researchers recommended the following action items be completed for the November 2008 Plenary: the development of a procedural document that outlines how the AMS Web site is maintained; ultimately this document is to be voted on by the organization and then implemented. A second action item was to appraise what content on the AMS Web site should be preserved, and the final action item was to research the best process/strategy for preserving the archival content of the AMS Web site and to propose recommendations to TEAM Canada.

These action items were completed in time for the November 2008 TEAM Canada Plenary Workshop. Concerns were raised that the AMS was still uncertain of which parts of the Web site it wished to preserve and why. Four key questions that needed to be answered were identified as being: 1) what to capture, 2) how often to capture, 3) how much to capture, and 4) how long to preserve what is captured.

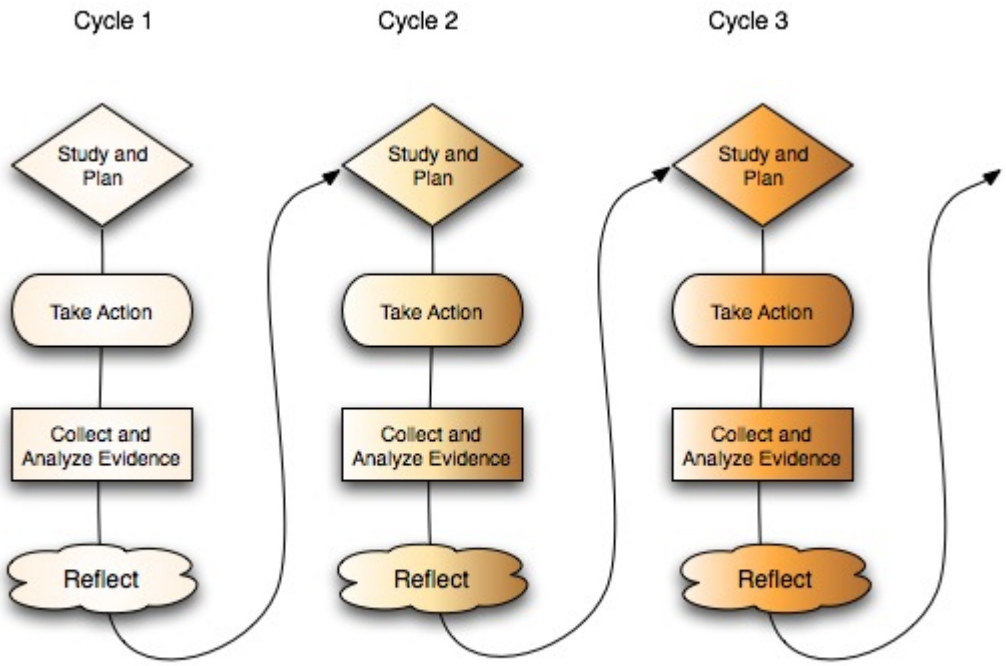
Three further action items were assigned to the Graduate Research Assistants at the November 2008 Plenary: conduct a (re)appraisal of the AMS Web site content, based on further clarification of what material the AMS wishes to preserve and why; identify the technological option(s) that meet the AMS's appraisal objectives and its technological, financial and human resource constraints; and identify the on-going costs of implementing the identified technological

options. These three items were to be completed by March 2009 and presented to TEAM Canada at the May 2009 Plenary. A summary of findings is included in this report.

At the May 2009 Plenary it was decided that enough data had been collected for the AMS organization and that this final report be written to reflect the several possible solutions articulated and to build an Action Plan that includes strategy, protocols, functional requirements, procedures and expected outcomes.

It is recommended that the AMS implement the Action Plan with the assistance of InterPARES researchers to allow the researchers to test the plan and to reflect on the results.

The Center for Collaborative Research highlights the importance of progressive problem solving using the action research method:[3]



Progressive Problem Solving with Action Research

Involving InterPARES researchers in the implementation process will ensure that the AMS receives a plan that is beneficial to them as well as InterPARES developing an understanding of how the plan will transcend organizations, person, or community. The researchers can find out how well the recommended action plan serves the AMS and suggest the

[3] Diagram from the Center for Collaborative Research Web site. Available at: http://cadres.pepperdine.edu/ccar/define.html.

distribution, translation and teaching of the plan to other organizations. InterPARES involvement in implementation will ensure a satisfactory result for all stakeholders involved.

C. Description of Context:

Provenancial

The AMS is a society for students, run by students, located on the UBC Point Grey campus, a not-for-profit private institution. The AMS is committed to the promotion of high-quality student learning. It advocates students' interests, as well as those of UBC and post-secondary education as a whole. AMS members are comprised of all UBC students who pay fees, as well as students at colleges affiliated with UBC such as Regent College and the Vancouver School of Theology.

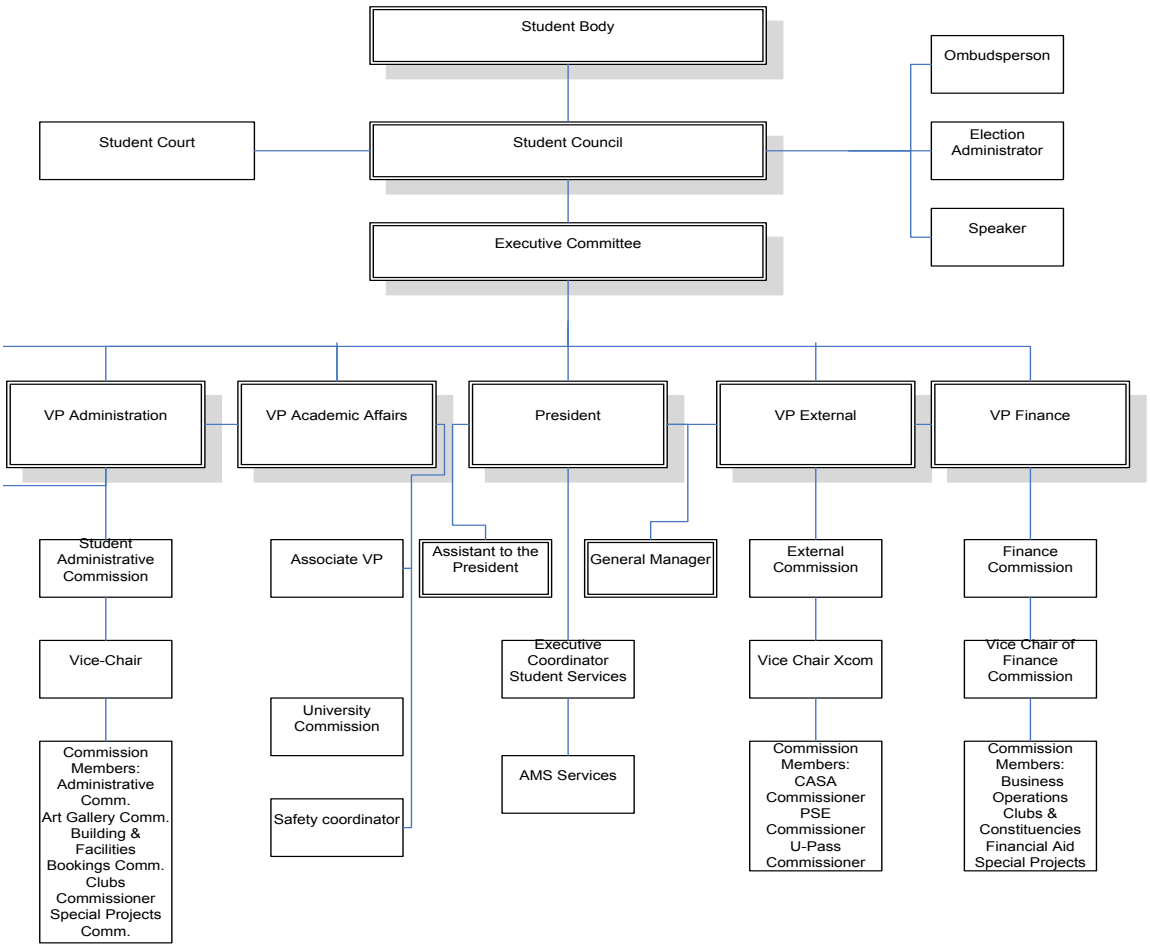
It states its mission to be to "improve the quality of the educational, social, and personal lives of the students of UBC."[4] Additionally, the Society seeks to provide its members with diverse opportunities to become exceptional leaders. The AMS's priorities are determined by its members. The society fosters communication, both internally and externally, to be democratic, fair, accountable, and accessible to its members. It provides services students want and can use. The AMS seeks to engage students in campus life and to empower students to further the goals they set for themselves.

The AMS is governed by a forty-five member Student Council. Council members consist of elected representatives from the various faculties/student constituency groups of the Society, and are elected annually by the Societies members. Specifically, Council consists of the President and Vice-President; the Directors of Administration and Finance; the Coordinator of External Affairs; representatives from the various undergraduate and graduate societies and schools; student representatives on the UBC Board of Governors and Senate; representatives of the Graduate Student Society; and the AMS Ombudsperson. The President and Vice-President; the Directors of Administration and Finance; the Coordinator of External Affairs comprise the five-member Executive Committee, which is a separately-elected part of Student Council

[4] See the Mission statement of the AMS as published on its Web site: www2.ams.ubc.ca/index.php/ams/subpage/category/about_the_ams.

responsible for directing the overall operations of the AMS.[5] The Executive Committee is chosen through campus-wide elections in which all AMS members may vote.

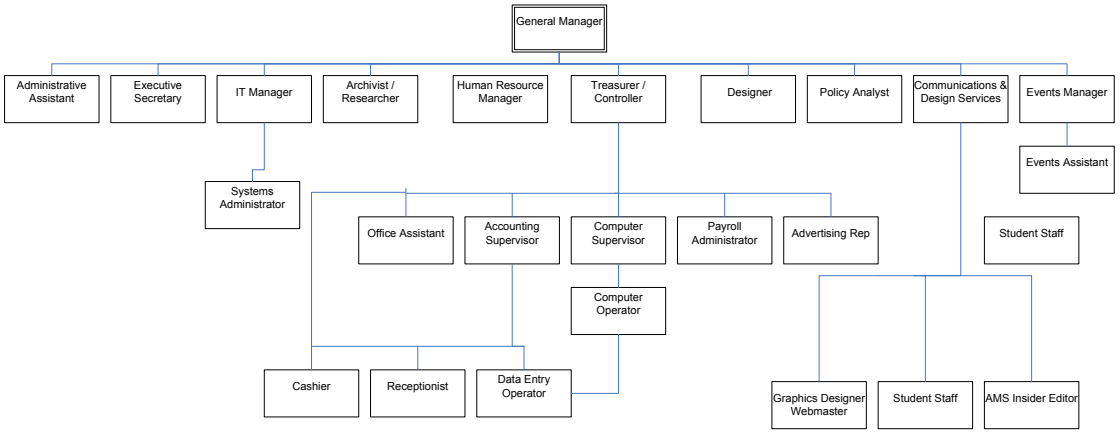
The overall structure of the AMS shows the organization of the Society as a whole:



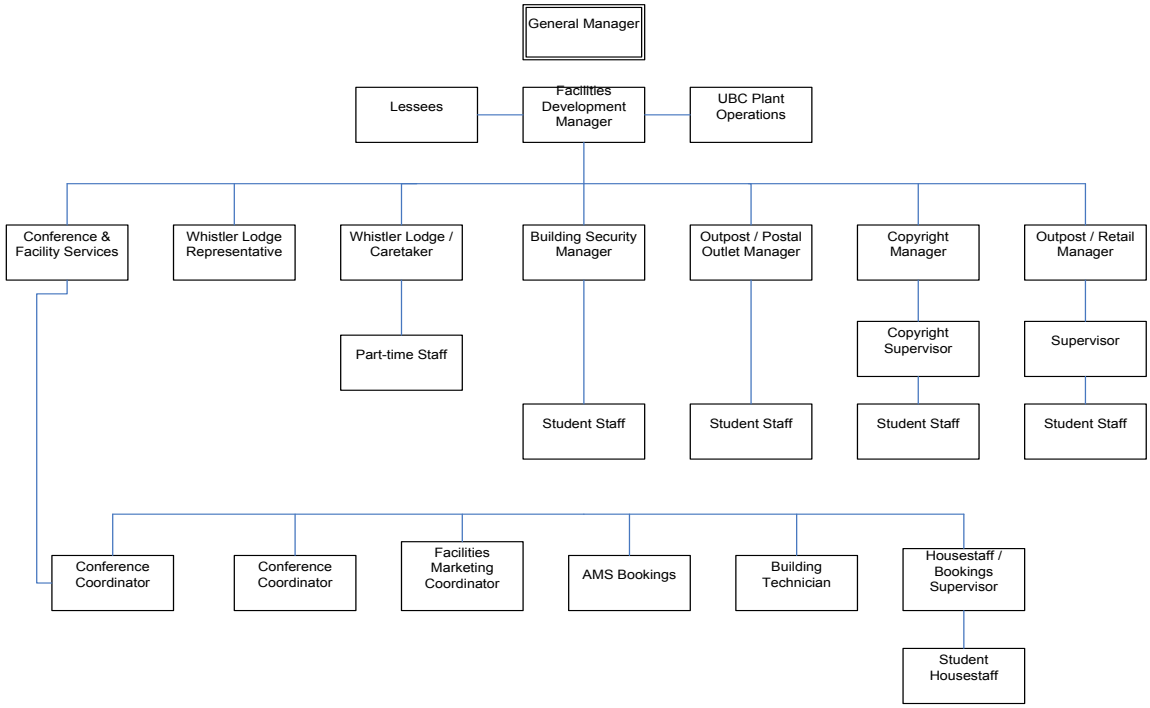
Elections are held every January. Due to annual elections there is a high rate of turnover in upper management at the AMS. Operational continuity is provided through an extensive archival record that is maintained by each outgoing student executive administration and by the presence of permanent, full-time, non-student support staff members. Full time staff members include a General Manager, an Administrative Assistant, an Executive Secretary, an Information Technology Manager, a Researcher/Archivist, a Treasurer, a Designer, a Policy Analyst, a

[5] These are the official titles of the Executive Committee; in practice, however, all members of the committee (with the exception of the President) are known as Vice-Presidents: Vice President Academic, Vice-President External, Vice-President Administration, and Vice-President Finance.

Communications Manager, and an Events Manager. These support staff members oversee much of the AMS operations, shown in the organizational charts listed below.



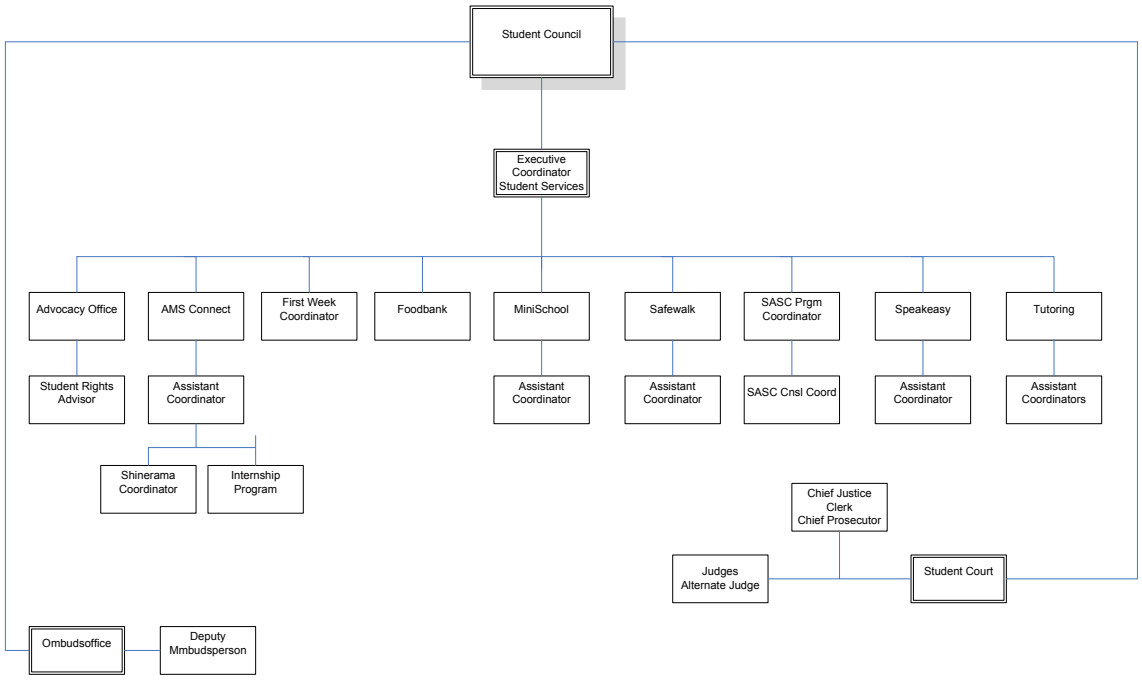
The General Manager also oversees the many AMS business operations:



Several commissions oversee specific aspects of the AMS's s operations. Commissions include, but are not limited to the Student Administrative Commission (SAC), the External Commission (XComm), and the Finance Commission. Generally, commissions oversee the administration of student clubs and the administration of the Society's external affairs, its

relationship with the university, the surrounding community, and the provincial and federal governments. Commissions are generally run by student executives, but commissioners are prohibited from being members of Student Council themselves.

The Student Council oversees the many services that the Society runs for students:



The AMS subscribes to a tacitly understood code of ethics when conducting its regular business affairs. The Society adheres to its own written codes, policies, and procedures; these outline the Society's day-to-day administration as well as define the roles and responsibilities of its staff and members. These documents include a Code of Procedure, Bylaws, Policy Manual, and Executive Procedures Manual. Additionally, with regard to administrative records, records in the custody of the Society are generally not disclosed without receiving prior consent from the Society's Archivist (also the Privacy Officer) and/or the General Manager.

The AMS Archives is a combined archives and records centre for the Alma Mater Society of the University of British Columbia. It houses semi-active and inactive records of the AMS, mostly to be used by staff within the organization, but the records are also available for consultation by the public.

Juridical-administrative

The AMS subscribes to a tacitly understood code of ethics when conducting its regular business affairs. The Society also abides by its own Code of Procedures and Constitution, and is governed by self-regulating bylaws.[6] These documents set forth the behavioural and regulatory rules for how the AMS operates as an organizational whole; including outlining how the Society conducts its day-to-day administration as well as defining roles and responsibilities of its staff and members. However, there is no external regulatory body that the AMS has to adhere to, and there will be no legislative penalty for the AMS if such internal directives are not followed.

The AMS is a non-profit Society incorporated under the *Society Act*. Incorporation under this act allows the AMS to act as a "natural person of full capacity" to carry out its business in pursuit of its stated purposes.[7] Further, the *Act* regulates the Alma Mater Society to a limited extent. With regard to governance and financial affairs, the *Act* outlines the legal rights, responsibilities, and obligations that the AMS has regarding its finances, property, members, and directors. For example, certain financial records as well as a compulsory annual audit statement must be made available to members of the Society upon request and within a reasonable amount of time.

Although the AMS is a separate entity from the University of British Columbia, it is to some degree, affected by legislation governing the University. A student society, as defined by the BC *University Act*, is an organization incorporated as a society under the *Society Act* whose purpose is to represent the interests of the general undergraduate and/or graduate student body (this definition excludes national and provincial student organizations). Under the *University Act*, the UBC Board of Governors has the authority to collect student society fees and is required to remit them to the AMS in a timely manner.[8]

The AMS is subject to specific laws such as copyright legislation and privacy legislation. Because the AMS collects and maintains personal information from its members, employees, and others, it is subject to the B.C. *Personal Information Protection Act* (PIPA). According to the internal policy, *AMS Personal Information Protection Policy*, the AMS is fully committed to complying with this legislation: "We will inform our employees, volunteers, members, suppliers,

[6] AMS Constitution: www2.ams.ubc.ca/images/uploads/AMS_CONSTITUTION_NEW_2008.pdf; AMS Code of Procedures: www2.ams.ubc.ca/images/uploads/New_Code_2008-_updatedMay_formatted.pdf; AMS Bylaws: www2.ams.ubc.ca/images/uploads/AMS_Bylaws_NEW_2008.pdf.
[7] See the *BC Society Act*, 4 (1)(d), available at www.qp.gov.bc.ca/statreg/stat/S/96433_01.htm.
[8] See the *BC University Act*, available at www.qp.gov.bc.ca/statreg/stat/U/96468_01.htm.

and customers of why and how we collect, use and disclose their personal information, obtaining their consent where required, and only handle their personal information in a manner that a reasonable person would consider appropriate in the circumstances."[9] The AMS does not collect any personal information without fully disclosing the reasons for which this information is to be collected and used, and likewise will not use or disclose this information for any other reason.

It is the responsibility of the AMS under PIPA to ensure the security and confidentiality of all personal information in its possession. Confidential materials as defined by the PIPA legislation are always printed out and segregated from the general records in the archives. With this in mind and with regard to the AMS Web site, the AMS must make absolutely certain that no personal information appears on the Web site at any time. Failure to do so would make the AMS liable under the PIPA legislation.

There is no legislation that directly affects Web content and creation. However, as seen above, PIPA legislation governs content of the site to the extent that no personal information can be publicly displayed on the Web site. Copyright legislation impacts Web content in that administrators need to ensure that all content published online by the AMS is either original work or outside of the scope of copyright requirements.

Procedural

Each department is responsible for the creation and management of its own records. Most records are managed in an *ad hoc* fashion. The Archivist is in charge of facilitating the long-term preservation and access to the semi-active and archives of the AMS. However, there is no Records Management Policy in existence; instead the Archivist relies on more informal means such as "friendly reminders" to staff.

Student executives are required to submit their records to the archives every year at turnover that follows the annual student elections; a practice that is generally followed.[10]

Records are transferred to the AMS Archives in an *ad hoc* manner. Records in the custody of the Archives include written, aural and photographic materials relating to all aspects

[9] See the *AMS Personal Information Protection Policy*, available at www.amsubc.ca/index.php/ams/subpage/category/privacy_policy.

[10] See InterPARES 3 Project, TEAM Canada, "Case Study 09 – Alma Mater Society of the University of British Columbia – Web Site Preservation: Workshop 03 Action Item 21 – Reappraisal of AMS Web site Content," pp. 2-3 (www.interpares.org/ip3/display_file.cfm?doc=ip3_canada_cs09_wks03_action_21_v1-3.pdf), in which the GRAs describe a scenario where records thought to exist elsewhere were in fact only stored on the AMS Web site.

of the AMS's s mandate. At present, the vast bulk of the archival holdings are maintained in the original format.

The AMS Archives has embarked on digitization projects that have digitized Council and SAC minutes dating back to the 1980s and are in the process of a photo project that will make the photo collection accessible digitally.

Although most digital records are maintained on the creators' computer systems, there is a formal system of capturing Executive and senior management emails that transfers copies of these into the AMS Archives. Active records remain in their creators' offices, and the AMS maintains a server on which all digital files are to be stored and backed-up regularly. Many records creators print and physically store important born-digital records, retaining these in their offices.

Documentary

The AMS Web site is linked to the fonds of the Alma Mater Society of the University of British Columbia. Technically, as no records are produced by the Web site, it is not a part of the fonds, but of its dissemination materials. However, if the Web site as a whole is to be judged as a record of AMS activity, then it could be considered a part of the Alma Mater Society fonds.

Certain components of the Web site may be related to other records elsewhere in the organization, but these components are not explicitly linked by an archival bond. For example, job postings or volunteer opportunities published on the Web site may also exist in hardcopy or basic copy in the offices of human resources or other staff; similarly, the news and events blogs on the Web site may discuss events that independently generated records elsewhere in the organization, but the Web content and the records themselves are not necessarily linked in any formal or explicit manner.

The archives have copies of previous versions of the AMS Web site in its custody. These were acquired on an informal basis, as the archivist appraised various portions of the Web site as potentially having long-term value and printed out these portions (mostly job postings and events calendars). These paper print-outs are filed and preserved with the rest of the Society's hardcopy archives.

Technological

The AMS Web site currently operates on a proprietary server-based system protected by a firewall, although the organization is in the process of moving the Web site to its own internal server that it rents from UBC. The AMS operates a solely Windows based platform.

Many types of media are created by the AMS for the Web site, including: textual, audio, video, digital images, photographs, and digital documents, although there are still no guidelines in effect concerning the creation of these media types.

Throughout the course of its activities, the AMS creates records in multiple formats, although there is no standardization. Examples include .pdf, .doc, .jpg, .xls, .mov and .gif. Some projects require the creation of the same document or record in various formats.

The Web site runs PHP to pull data out of a MYSQL Database and formats and presents this data "on the fly" to users as navigable Web pages. Two servers are currently used for the AMS site. The Web site as a whole consists of an elaborate series of inter-connected Web pages that represent the various branches, departments and associated functions of the Society. In general, the Web site reflects the ongoing activities of the AMS. The Web pages are not necessarily by-products of the Society's activities, but rather are updated periodically to reflect and publicize the Society's actions and raise its profile in the campus community.

There are no written policies or procedures that govern how or when Web content is updated. The InterPARES 3 TEAM Canada researchers produced a document that governed updating of Web site content that was ultimately to be voted on and implemented. However, the AMS deemed it unnecessary to put into practice and therefore, Web site content is still ad hoc in its creation and management. Changes in Web site content may be initiated by up to forty different users, however, the upload procedure has been streamlined; instead of allowing these individuals to upload their own Web content, all changes/new content is now funnelled through two staff—the Communications Manager, who passes approved changes/content on to the Web Editor for upload.

IT resources are limited at the AMS, with one IT Manager supervising the organization's entire technological capital. Therefore, any preservation strategy for the Society's Web site had to be easily implemented and straightforward enough to easily teach to a largely student staff complement with a diverse range of technological skills and knowledge.

D. Narrative answers to the records case studies questions for researchers

The AMS Web site is created to disseminate information. Primarily, the Web site informs members of the Society about the services, resources, and opportunities that it provides; however, it is also useful in informing the larger campus community as well as the public at large about news, events and issues affecting the student population at UBC.

The Web site consists of blogs, events calendar, job postings, news postings, and other information resources for students, such as tutoring information and other services. Each individual organization within the AMS is responsible for its own Web content. The Web site is dynamic and constantly being changed and updated as priorities change, events are planned and carried out, elections occur and so on. Content for publication is related to the ongoing business activities of the AMS as a whole; however, the Web site itself is not used for recordkeeping purposes nor does it contain or generate official records as a by-product.

Student services, businesses, members of student government and all branches of the AMS submit content for publication on the AMS Web site to the Web Editor (a member of student staff). The Web Editor "posts" content to the public Web site by copying and pasting files into the content management system (CMS). (The CMS used for the Web site is Expression Engine, an application that runs on a Web browser and allows for intuitive and user-friendly Web editing).

To ensure accuracy, reliability and authenticity, substantive edits are subject to prior approval by the Communications and Design Services Manager; however, most edits are small or incremental in nature and thus do not necessitate prior clearance. The editing process consists mostly of proofing for grammatical or similar errors, as the students who create the various Web pages are responsible for articulating the intellectual content of those pages. The Communications and Design Services Manager is usually copied on any e-mail that contains the requested content update; however, if she is not included on the e-mail, the Web Editor may forward the requested edit to her in advance of making the changes. However, since changes to the Web site are requested often and usually occur in small increments, the Web Editor may simply post the content without receiving (or needing) prior approval. There is no process in place that ensures Personal Information Protection Act legislation is followed other than relying on the Communications Manager and Web Editor's sound judgement. Changes to the Web site content are not formally recorded.

The Web site consists mainly of text and images, although QuickTime videos and other digital media formats have been published on the site in the past. If and when video content appears on the Web site, it does not actually reside on the AMS servers, but is linked from servers elsewhere on the Internet via an embedded URL link.

Maintenance of each page's graphical user interface/aesthetic consistency is largely automated through the formatting templates designed and maintained by Whitematter. As content is copied into the CMS, most of the formatting of the Web site occurs automatically according to the type or category of the content uploaded, as identified by the Web Editor at the time of update.

The digital components of the AMS Web site include graphical and textual components created in .pdf, .doc, .jpg, and .gif. The software used to create and update the Web site content is Expression Engine, which uses a PHP content management system to pull data out of an MYSQL database and format and present this data "on the fly" to users as navigable HTML Web pages.

No metadata are manually added to any of the files or digital components of the Web site by content creators or administrators. Individual file formats and software systems may automatically generate metadata internally within the digital components themselves; these allow digital components and/or files to communicate data about themselves to software programs or to other computers or peripherals.

Digital components of the Web site are stored in multiple locations: a copy is retained on the Whitematter server (this is the copy viewed by users of the public Web site, but is in the process of being moved to the AMS's own in-house server); a basic copy is sometimes kept on the Web Editor's computer; a basic copy is retained temporarily in the Web Editor's (and sometimes the Communications and Design Manager's) e-mail inbox; and basic copies may be retained by the requestor of the change and/or in the requestor's e-mail outbox. There are no formal procedures for retaining or handling redundant copies of each change to Web content.

A lack of formal procedures could result in personal information guarded by the PIPA legislation inadvertently ending up published on the Web site. Additionally, records thought to exist elsewhere have only existed as Web content and lost as a result of the frequency of change to Web content.

Although the AMS wishes to preserve its Web site for informational purposes only, regardless of whether there are records contained within the site, the TEAM Canada researchers decided to continue with research into Web site capture tools.

E. Narrative answers to the applicable Project research questions

Using the AMS case study of preserving their Web site as a basis, the Graduate Research Assistants attempted to answer a variety of the general project research questions[11] in their report writing. Namely, how can we adapt the existing knowledge about digital records preservation to the needs and circumstances of small and medium sized archival organizations or programs? What are the nature and the characteristics of the relationship that each of these archives or programs should establish with the creators of the records for which it is responsible? What knowledge and skills are required for those who must devise policies, procedures and action plans for the preservation of digital records in small and medium sized archival organizations or programs? What action plans may be devised for the long-term preservation of these bodies of records? Can the action plan chosen for a given body of records be valid for another body of records of the same type, produced and preserved by the same kind of organization, person, or community in the same country?

How can we adapt the existing knowledge about digital records preservation to the needs and circumstances of small and medium sized archival organizations or programs?

We sought, with our research, to identify methods for Web site preservation that had been successfully implemented in other similar organizations, as well as look to large organizations to learn from their knowledge. We also investigated methods that had not been currently implemented. Many of the large organizations have been instrumental in developing methods for Web site capture and preservation and we looked to these organizations for tried and tested methodologies. Among the most useful large organizations currently preserving Web sites were the Library of Congress, the Internet Archive, the National Archives UK, and the National Archives of Australia. Each of these institutions was helpful in developing our understanding of what components were necessary to be included in a preservation strategy. Much of the information is easily adaptable to the needs of small and medium sized archival organizations or programs, and without this research many smaller institutions would not be able to undertake such preservation programs. The Internet Archive has been developing open source solutions for remote harvesting operations that do not require a monetary output, but do require fairly extensive technological knowledge. The National Archives of the United Kingdom have

[11] See www.interpares.org/ip3/ip3_questions.cfm.

conducted research into best storage medium, a simple guide to Archiving Web sites, as well as researching optimum file formats for data creation. The National Archives of Australia has produced research on metadata requirements that are key to effectively managing all digital records, including records of Web-based activity. They have also researched solutions for recording evidence of Web-based records on frequently changing Web sites when infrequent crawls are in place. The Library of Congress has also conducted research into metadata specifically for preservation (PREMIS A data dictionary and supporting XML schemas for core preservation metadata needed to support the long-term preservation of digital materials) as well as developing other metadata schema (METS (Metadata Encoding and Transmission Standard) A metadata structure for encoding descriptive, administrative, and structural metadata that produces Encoded Archival Descriptive Finding Aids).

What are the nature and the characteristics of the relationship that each of these archives or programs should establish with the creators of the records for which it is responsible?

We established that many of the tasks of the archivist looking to implement a program of Web site capture and storage would be made easier by having the cooperation of the creator of the Web site. This is possible in the case of the Alma Mater Society as the only Web site they wish to preserve is their own, and therefore, has the capacity to make certain requests to the Web site creators. We established a variety of components that could benefit those tasked with Web site preservation; namely, uploading content in specific file formats and the addition of metadata to Web page headers. If many documents now uploaded to the AMS Web site as .doc, .xls, and .ppt were all converted to .pdf files before upload it would allow for the need to only preserve a single file format and allow access to both PC and MAC users. If preservation metadata were added to the Web page templates, the viability, renderability, understandability, authenticity, and identity of digital objects in a preservation context would be preserved. Not all archival institutions have the luxury of dictating to Web site owners' elements of Web creation, but these are valid requirements if the institution is able to make requests of the Web site creators.

What knowledge and skills are required for those who must devise policies, procedures and action plans for the preservation of digital records in small and medium sized archival organizations or programs?

The AMS approached the InterPARES team with a view to devise strategies for preserving their digital records with the caveat that due to high turnover and limited resources, any solution must be simple, cost-effective and taught easily to in-coming student staff. For each capture and storage solution researched this caveat was kept in mind. Findings were presented to the AMS archivist as well as TEAM Canada researchers that suggested the level of complexity for each option discussed. It is apparent, however, that those devising policies, procedures and action plans for the preservation of digital records must have a relatively high level of technical skills and a basic understanding of the terminology and methodology involved with digital records preservation. It is of critical importance to have sufficient knowledge of the technology to either prepare effective specifications for use by a third party (be it an in-house technology department or an organization that is used to outsource the preservation process), or to undertake the work oneself.

What action plans may be devised for the long-term preservation of these bodies of records?

There is no single, definitive solution to be applied to Web site archiving. Strategies will depend upon a variety of factors including the presence (or absence) of records on the site, content ownership, technical capability, costs and storage abilities. Therefore, there are several action plans that could be devised for the long-term preservation of an institutional Web site. The action plans range from extremely technical solutions that are highly effective and address the dynamism of certain back-end database driven Web sites to simple, relatively inexpensive solutions that preserve a snapshot of the Web site in time. Tools are available that facilitate Web site archiving. The tool chosen will depend greatly on how much information the archiving organization wishes to preserve, the technical abilities of staff, and a thorough risk assessment. An approach that is based on good management practices and begun as early as possible in the lifecycle of the digital resource will be effective at least for the short to medium term.

There are many considerations for an organization about to embark on a Web site preservation program. Factors include technical ability; rights management; training; resource description, documentation and access; choice of file formats; validation checks; disaster

recovery planning; storage medium, standards, and which method for Web site capture is the most effective.

Technical Ability

As previously stated, it is important for anyone involved in the preservation of digital materials to have some understanding of what is involved. The individual responsible does not have to be a computer scientist, but must be knowledgeable enough to have an informed exchange with those involved in the preservation strategy as well as being able to set forth realistic requirements to a third party. Some strategies require an intensive knowledge of the technological environment in order for it to be implemented; while others require a minimal amount of knowledge to implement and succeed. Web sites that comprise static documents and incorporate little or no interactivity are relatively simple to deal with. However, sites that incorporate high levels of interactivity and comprise dynamically generated pages are very complex and prove more difficult to archive effectively.

Policy / Recordkeeping Requirements

Policies, procedures and criteria for a Web site archiving program are critical in the emerging digital environment. They ensure that the aims and objectives of the institution are carefully considered and reviewed; that collections development supports the institutional mission and priorities; and ensure accountability to the funding agencies and the wider academic community. Elements to consider including in a policy are: a policy statement, the goals and objectives of the policy, related documents and or legislation, scope of the policy, persons responsible for policy implementation, scope of collections, coverage, an outline of digital resource types accepted, rejection criteria, evaluation criteria, viability, and collection levels. These may be broken up into more than one policy.

Recordkeeping: All data associated with the archiving of Web sites should be included in retention schedules that govern the institution's records. Web pages should be subject to the same records management controls as other electronic records, since they provide evidence of the online activities of the organization. In addition to improved records management, the organization would benefit in terms of costs associated with storage if effective disposition schedules were in place. To ensure long-tem accessibility of data it is essential that storage media is refreshed on a regular basis. If the organization stores each iteration of the Web site

indefinitely then the costs associated with refreshing media will soar over time as the data collected grows.[12]

Metadata

Metadata is the key to effectively managing all records, including records of Web-based activity. Ross Harvey, Library and Archives Professor and preservation expert, asserts that "Preservation metadata is now considered an integral part of the strategies required for long-term maintenance of and access to digital materials…"[13] The Australian *Guidelines for Archiving Web Resources* describes suggested metadata requirements for different scenarios:

For individual records on Web sites and for other records of Web-based activity, this means using metadata to describe:

- Date and time of creation and registration of the record into a recordkeeping system;
- Organizational context;
- Original data format;
- The use made of the record over time, including its placement on a Web site;
- Mandates governing the creation, retention and disposal of the records; and
- Management history of the record following creation – including sentencing, preservation and disposal.

For copies or snapshots of entire collections of Web resources, metadata should include:

- Date and time of capture;
- Links to the universal resource indicator (URI) including information about version and date of link to specified URI;[14]
- Technical details about the Web site design;
- Details about the software used to create the Web resources;
- Details about the applications (including search engines) that supplement the Web resources; and
- Details about the client software needed for viewing the Web resources[15]

[12] The information set forth in this section is extremely basic. To learn more about electronic recordkeeping requirements please see: McLeod, Julie and Catherine Hare, eds. *Managing Electronic Records* (London, UK: Facet Publishing, 2005); Erlandsson, Alf, *Electronic Records Management: A Literature Review*, ICA Study 10 (Paris: International Council on Archives, 1997), available at: www.ica.org/sites/default/files/10litrev_1.pdf; Evans, Joanne, Sue McKemmish and Karuna Bhoday (2006), "Create Once Use Many Times: The Clever Use of Recordkeeping Metadata for Multiple Archival Purposes," *Archival Science* 5: 17-42; ICA Committee on Electronic Records, *Guide for Managing Electronic Records From and Archival Perspective* (Paris: International Council on Archives, 1997: and ICA Committee on Current Records in the Electronic Environment, *Electronic Records: A Workbook for Archivists* (Paris: International Council on Archives, 2005), available at: www.ica.org/sites/default/files/Study16ENG_5_2.pdf.

[13] Harvey, Ross. *Preserving Digital Materials* (Munich: K. G. Saur, 2005), 83.

[14] The Australian *Guidelines for Archiving Web Resources* distinguish between a URI, URL, and URN thus: Universal resource indicator (URI) a general purpose namespace mechanism; Universal resource locator (URL) an instance of URI that is the address of some resource, accessible by means of a protocol such as HTTP; Universal resource name (URN) an instance of URI that, unlike a fragile URL, is guaranteed to remain available (Jon Udell, *Practical Internet Groupware* (Sebastapol, CA: O'Reilly, 1999), 471.)

It is recommended that a metadata audit be performed when embarking on a Web site archiving and preservation program. This will ensure that captured resources have sufficient metadata attached to effectively preserve the accuracy, authenticity, reliability, accessibility and disposition of the resource and allow access and preservation activities to occur.

Rights Management / Intellectual Property Rights

Issues surrounding intellectual property rights, such as copyright concerns and moral rights have a substantial impact on any digital preservation process and this is no different for Web site preservation and archiving. Maggie Jones and Neil Beagrie argue that "The intellectual property rights issues in digital materials are … more complex and significant than for traditional media and if not addressed can impede or even prevent preservation activities."[16] Jones and Beagrie justify their argument by suggesting that not only content, but any associated software may be subject to intellectual property rights, and warn that, "Simply copying (refreshing) digital materials onto another medium, encapsulating content and software for emulation, or migrating content to new hardware and software, all involve activities that can infringe intellectual property rights unless statutory exemptions exist or specific permissions have been obtained from rights holders." Due to the nature of digital materials, strategies for continuing preservation and access may necessitate the migration of the materials into new forms or an emulation of the original operating environment. Such activities may require permissions from rights holders to legally undertake such strategies.

A specific area that could potentially become problematic is in the area of Copyright Law. According to the Canadian Heritage Information Network (CHIN), "Copyright protects the expression of ideas that are fixed in any form of media."[17] This includes various Web site components, such as images appearing on a given site and the underlying software programming code:

> Copyright protects the majority of creations including, literary, dramatic, musical and artistic works, sound recordings and audio-visual works. Photographs are considered artistic works. Computer software programs including underlying code have been identified as literary works and they are therefore also protected by copyright. Except where works are created in the course of employment in the

[15] "*Archiving Web Resources: Guidelines for Keeping Records of Web-based Activity in the Commonwealth Government,*" from the National Archives of Australia, p. 17-18.
[16] Jones, Maggie and Neil Beagrie, *Preservation Management of Digital Materials. A Handbook* (London, UK: The British Library, 2001), 32.
[17] Pantalony, Rina Elster, *Protecting your Interests: a legal guide to negotiating Web site development and virtual Exhibition Agreements* (Ottawa, Canada: Minister of Public Works and Governments Services Canada, 1999), 13.

course of an employee's duties or where copyright has been assigned in writing to someone else, the author of the work is the copyright holder.[18]

Copyright holders should be established and permissions granted before embarking on a Web site preservation program.

Staff Development and Training

Carefully designed staff training and continuous professional development can play a key role in successfully managing any digital preservation program. All those responsible for digital preservations must have a degree of knowledge on the topic. Staff development and training can range from keeping up to date with the literature and new developments to participating in workshops and training modules put on by various institutions and organizations such as Archival Societies and educational institutions.[19]

Resource Description, Documentation and Access

Some form of classification description is essential in order to manage any archival collection and make it accessible to users; this is no different for digital collections. Major cataloguing standards, such as MARC 21 and ISAD(G), have been successfully applied to the description of archived Web sites. Cataloguing and classifying archived materials allows user access to them.

Resources should be supplied with appropriate and sufficient documentation to satisfy the requirements for informed use by members of the research community. The documentation should relate to both the content and the technical format of the resource. Documentation should also provide information about the context in which resources were created and maintained before archiving, and about the relationships between the digital resource and other information sources.

Disaster Recovery Planning

The development of a disaster recovery plan that is based on sound principles, has buy-in from management and can be activated by trained staff will greatly reduce the severity of the impact of disasters. The plan will need to address the restoration of both the content of the

[18] Ibid.
[19] The Society of American Archivists is one institution that organizes many workshops and Web seminars. For a calendar of current opportunities see http://saa.archivists.org/Scripts/4Disapi.dll/4DCGI/events/ConferenceList.html?Action=GetEvents.

archive, and the technical and operational infrastructure required to support it. Elements to be included in a plan should be:

- Ensure staff are trained in counter disaster procedures;
- Create archives copies of data resources each time a collection of materials takes place; Store archived copies on multiple media;
- Store archived copies on and off site;
- Complete documentation of the hardware and software infrastructure as well as operating procedures and manuals;
- Copies of all software required to operate the systems.

It is also important to test the plan to discover any issues that may have been overlooked before the event of a disaster occurs. This is also helpful to staff to allow them to become familiar with the procedures before hand. As with most policies, it is recommended that the disaster recovery plan be revisited as systems and circumstances change.

Validation Checks

Once the Web site has been captured and transferred to the institution's archival environment, checks must be conducted to ensure that all the parts of the Web site captured are working as they should. Checks include, but are not limited to: manually going through and clicking on all the hyperlinks; randomly clicking on links; or employing the use of a link testing application to help automate the checking process by testing to see that all links are working,[20] checking that the files can be read, checking files for completeness and accuracy and checking functionality within the files. Checks should be carried out whenever a Web site archive has taken place to ensure the content and structures of the deposited data resources are intact.

File Formats

With any Web site preservation program (like any digital preservation program) it is recommended that accepted file formats are defined before embarking on any collection strategy. The adoption of a single file format ensures that sustainability costs are minimized when a file format of choice is built into the records creation process.

According to Evelyn Peters McLellan, InterPARES 2 Co-Investigator, "it has become common practice for digital records repositories, including archives, to accept certain digital file

[20] See, for example: Link Checker Pro: www.link-checker-pro.com/; Site Audit: www.blossom.com/site_audit.html; Cyber Spyder Link Test: www.cyberspyder.com/cslnkts1.html; Link Sleuth: http://home.snafu.de/tilman/xenulink.html.

formats for long-term preservation while rejecting others."[21] In her report, McLellan surveyed institutions to gather data regarding file format specifications. Her research showed that there was a plethora of definitions, acceptable/unacceptable formats, and preservation initiatives for file formats. The PREMIS *Data Dictionary for Preservation Metadata* gives the most useful definition: "a specific, pre-established structure for the organization of a digital file or bitstream." McLellan notes that "This pre-established structure includes how the data are encoded, which is the way in which the bits are interpreted to produce text, images and sound."[22] This is important to understand as it highlights why it is essential to specify acceptable file formats to a specific repository. McLellan goes on, "Some types of encoding are synonymous with specific file formats; for example, MP3 encoding is used to encode the MP3 File format."[23] This is simple enough to understand, but it gets increasingly complicated. Take plain text files for example, McLellan points out that, "many formats can have different encodings: even a "plain text" file can be encoded as ASCII, EBCDIC or Unicode, all of which have a number of variants."[24] The plain text file has three different types of encoding, so obviously image and music files are much more complicated. McLellan explains, "Encoding can be problematic in audio and video file formats because the optimal encoding for storage and transmission often involves compression (removing bits from the digital files to reduce their size), which can often hinder preservation efforts."[25] McLellan notes further difficulties to the file format debate: "The encoding issue is further complicated by the fact that TIFF, WAVE, AVI and other common image and audiovisual formats are not file "formats" per se, but rather file "wrapper formats" (also called container formats), which are designed to combine multiple bitstreams into a single file."[26] Encoding, compression and bitstream combinations all complicate how file formats are preserved over the long-term. These are also reasons why many institutions call for open formats that are well documented to ensure that sufficient documentation is available to give the collecting institution a chance of preserving digital records for the long-term.

[21] Peters McLellan, Evelyn, "General Study 11 Final Report: Selecting Digital File Formats for Long-Term Preservation," InterPARES 2 Project (March 2007), 1. Available at www.interpares.org/ip2/display_file.cfm?doc=ip2_gs11_final_report_english.pdf.
[22] Ibid, 2.
[23] Ibid.
[24] Ibid.
[25] Ibid
[26] Ibid.

Adrian Brown of the National Archives of the United Kingdom has identified criteria to consider when selecting file formats for data creation. The criteria include:

- Ubiquity
- Support
- Disclosure
- Documentation quality
- Stability
- Ease of identification
- Intellectual property rights
- Metadata support
- Complexity
- Interoperability
- Viability
- Re-usability

Although the research does not recommend actual file types, these criteria are important to bear in mind when selecting file formats.[27]

It is important that the archiving organization develops policy that states the types of file formats that are acceptable to archive. By restricting the range of file formats that an institution agrees to receive and manage, the organization can be assured that the file formats it collects adhere to the criteria stated above and that they adhere to current standards. If "good" file formats are collected, the difficulties in preserving them will be minimized as well as costs reduced.

Ross Harvey highlights problems associated with the multiplicity of file formats in use: "Many formats are proprietary, that is, they are the property of an owner who, for commercial reasons, is not willing to provide access to documentation about them, and who may require a fee to be paid for their use."[28] This is a reason why most experts recommend file formats that adhere to open standards. This is also a reason why many file format registries have been developed. The registries exist to provide reliable and detailed information about file formats. Examples of file format registries include: PRONOM[29] and the Global Digital Format Registry.[30] In April 2009 the Global Digital Format Registry initiative joined forces with the UK National Archives' PRONOM registry initiative under a new name—the Unified Digital Formats Registry (UDFR). The UDFR will support the requirements and use cases compiled for GDFR and will be seeded with PRONOM's software and formats database.[31]

The collecting organization can help promote sound records creation by publicizing those file formats that are most likely to be sustainable over a period of time and by encouraging

[27] Adrian Brown, "Selecting File Formats." Available at www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf.
[28] Harvey, Ross, *Preserving Digital Materials* (Munich: K. G. Saur, 2005), 141.
[29] PRONOM is a file format registry established by the National Archives (UK) to provide and manage information about file formats and software applications used. The PRONOM Web site can be found at: www.nationalarchives.gov.uk/pronom.
[30] The Global Digital Format Registry was also developed to support digital preservation. www.gdfr.info/.
[31] The Unified Digital Formats Registry is available at: www.udfr.org/.

records creation using these particular formats. Another alternative is for the collecting institution to convert all digital materials archived to the file format of choice once the material is in the archives.

Storage Medium[32]

Whichever capturing method is used, the archived Web site needs to be preserved and stored on a relatively stable electronic digital medium. Currently, no electronic digital medium can be considered archival due to concerns regarding the relatively short and/or unproven life spans of such media and to concerns regarding technological obsolescence resulting from rapid changes in the technological environment. Storage hardware is being continually developed. Current "state of the art" medium may be obsolete in 5 years time and simply impossible to maintain in 20 years time. Electronic media are not as permanent as is often thought. Manufacturers may claim satisfyingly long lifetimes for their media[33] but practical experience suggests that a realistic figure for the life of a magnetic tape may be 15 years, and for a CD 20 years, all depending on original quality, storage, handling, and usage. And even if the media lifetime is longer, the hardware to read it may not be available. For many media, a small imperfection that appears after some time may make the whole medium unusable.[34] Therefore, whichever medium is chosen for storage will need to be periodically checked and/or refreshed to counteract data loss.[35]

A variety of factors affect the longevity of electronic media, including storage conditions, quality of the products used, and the composition of the products due to the availability of better materials over time. Therefore, it is difficult to predict longevity. The Canadian Conservation Institute has put together a table that provides estimates of predicted longevity for various media storage types.

[32] The information presented here is at the most basic level. In this report we present basic storage medium for storing electronic media. It is possible to create a repository for digital materials. If you require more information take a look at the ISO Standard: ISO 14721: 2003, more commonly known as the Open Archival Information Systems (OAIS) reference model and OCLC and NARA. "Trustworthy Repositories Audit & Certification: Criteria and Checklist" Version 1.0, 2007. Available at: www.crl.edu/PDF/trac.pdf.

[33] 1995 Kodak research on their writeable CDs, reported at www.cd-info.com/CDIC/Technology/CDR/Media/Kodak.html, quoted a lifetime of 217 years under specified conditions.

[34] Jim Liden Sean Martin, Richard Masters and Roderic Parker, "The large-scale archival storage of digital Objects," DPC Technology Watch Series Report 04-03, February 2005.

[35] See The National Archives of the UK's Digital Preservation Guidance Note: 2, "Selecting Storage Media for Digital Preservation," by Adrian Brown, Head of Digital Preservation Research, August 2008. Available at: www.nationalarchives.gov.uk/documents/selecting-storage-media.pdf (accessed September 29, 2008).

Predicted longevity of electronic media[36]

Media type	Predicted longevity
Magnetic disks	
Hard disks	2–5 years
Floppy diskettes	5–15 years
Magnetic tapes	
Digital	5–10 years
Analog	10–30 years
Optical discs	
CD-RW, DVD-RW, DVD+RW	5–10 years
CD-R (cyanine and azo dyes)	5–10 years
Audio CD, DVD movie	10–50 years
CD-R (phthalocyanine dye, silver metal layer)	10–50 years
DVD-R, DVD+R	10–50 years
CD-R (phthalocyanine dye, gold metal layer)	>100 years
Other optical discs	
MO, WORM, etc.	10–25 years?
Flash media	?

It is therefore recommended that the archived Web site be stored in several environments—for example, on a hard drive and on DVD-R—and stored in the archives to counteract these storage concerns and help assure long-term access to the stored data.

In determining what type of storage media to store digital materials a number of factors need to be considered. These factors include longevity, capacity, viability, obsolescence, cost

[36] Canadian Conservation Institute, *Electronic Media Collections Care for Small Museums and Archives.* Available at: www.cci-icc.gc.ca/headlines/elecmediacare/index_e.aspx (accessed April 30, 2009).

and sustainability, again documented by Adrian Brown at the National Archives of the United Kingdom.[37] Brown displays a scorecard comparing common media types:

Media	CD-R	DVD-R	Hard disk	Flash Memory Stick and Card	Linear Tape Open (LTO)
Longevity	3	3	2	1	3
Capacity	1	3	3	2	3
Viability	2	2	2	1	3
Obsolescence	1	2	2	2	2
Cost	3	3	1	3	3
Susceptibility	1	1	3	1	3
Total	**11**	**14**	**13**	**10**	**17**

According to this chart, the top two storage solutions are Linear Tape Open and DVD-R, with a hard drive option a close third. Brown advices:

> In situations where multiple copies of data are stored on separate media, it may be advantageous to use different media types for each copy, preferably using different base technologies (for example, magnetic and optical). This reduces the overall technology dependence of the stored data. Where the same type of media is used for multiple copies, different brands or batches should be used in each case in order to minimise the risks of data loss due to problems with specific manufacturers or batches.

Joe Iraci, of the Canadian Conservation Institute, has additional comments regarding the differences of storage media. With regard to using optical storage media for storage, Iraci states: "the type of disc chosen and how it is recorded greatly impact[s] longevity." He highlights that "digital tapes have short lifetimes and need to be migrated/refreshed every 5-10 years" warns that "hard drives are not for long-term storage and data needs to be moved to a new hard drive every 2 to 5 years" and reminds us to "stick with technologies that are in widespread use and

[37] The National Archives, "Digital Preservation Guidance Note 2: Selecting Storage Media for Long-Term Preservation," August 2008. Available at: www.nationalarchives.gov.uk/documents/selecting-storage-media.pdf.

avoid new technologies" such as "Blu-Ray, Holographic Storage [and] Flash Media." Iraci also points out that "With all digital media, backups are critical in order to avoid sudden loss of information."[38]

Research such as that conducted by Adrian Brown and the Canadian Conservation Institute is invaluable when deciding what media to choose for the storage of institutional electronic records. It is clear that a variety of media should be chosen and that even with correct storage and handling the medium should be checked and refreshed regularly.

Standards

A number of standards are related to Web site archiving. HTML and XML are core technologies recognized as standards in the form of W3C[39] recommendations. Two standards exist in the area of records management: **ISO 15489-1/2: 2001** sets standards for records management practice, **ISO 23081-1: 2006** sets standards for records management metadata.

ISO 14721: 2003 sets the standard for defining fundamental requirements for a digital preservation system. More commonly known as the Open Archival Information Systems (OAIS) reference model, its concepts and terminology have been widely adopted by an international audience. It forms the basis for the certification scheme for trusted digital repositories.

ISO 19005-1: 2005 or the PDF/A standard has addressed the need for open digital file formats. The standard is "a file format based on PDF, known as PDF/A, which provides a mechanism for representing electronic documents in a manner that preserves their visual appearance over time, independent of the tools and systems used for creating, storing or rending the files."[40]

Web site Capture Methods

Currently, there are three options available for capturing Web sites and two types of Web sites built. The types of Web sites are either static or dynamic. A static Web site is composed of a series of pre-existing Web pages, all of which are linked to from at least one other page. A dynamic Web site generates Web pages on-the-fly from smaller elements of content. Such

[38] E-mail from Joe Iraci to Randy Preston, May 20, 2009.
[39] W3C or the World Wide Web Consortium is an international consortium where Member organizations, a full-time staff, and the public work together to develop Web standards.
[40] ISO-19005-1 - Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A-1).

content can be housed in a database, drawn from external sources and inserted into a Web page, or generated by scripts that respond differently depending on such factors as the date or time the Web page is accessed. The methods for capture vary depending on how much information the collecting institution wishes to preserve. Information includes functionality, metadata and the degree of authenticity, reliability and accuracy the collecting institution wishes to preserve. The three options are: direct transfer, remote harvesting and Web site mirroring.

Direct Transfer: The only way to fully recreate a Web site in a preservation environment is through Direct Transfer of data. Direct transfer works by acquiring a copy of the data directly from the original source. This requires direct access to the host Web server. Direct transfer then involves copying the selected files from the server and transferring them to the collecting institution. To guarantee continued functionality minor adjustments may need to be made to the archived site.[41] To ensure that the archived Web site is as authentic as possible, a recreation of the technical environment in which the Web site resides will need to be implemented within the archival setting. This means that the database or content management system will need to be installed in the archival environment, together with the necessary Web server and search engine software. Direct transfer is the only method that takes into consideration the dynamic nature of a Web site and is the only way to preserve all possible forms of dynamically generated data. However, the implementation and support of such a method will require staff with appropriate technical skills be available to install and maintain the system.

Remote Harvesting: The remote harvesting solutions offers three alternatives: a straight forward automated crawl of the Web site, a "snapshot" crawl with additional logs kept by the archivist to back up the data mined in the snapshot, and outsourcing the process to a third party. We offer remote harvesting collection methods as alternatives with the caveat that such data collection methods do not capture the entirety of all Web page possibilities that could be generated by a user request, if the Web site identified for capture is a dynamic site with an underlying back-end database used to house information generated on the fly. Also, using this method may result in the presence of broken links within the copied data environment as pages

[41] For example: The hyperlinks within the archived site may need to be adjusted from absolute links to relative links; and the appropriate search engine (the one used in the original environment) must be installed in the new environment to ensure that search functionality is preserved. For a more comprehensive explanation please see: Brown, Adrian, *Archiving Web sites* (London: Facet Publishing, 2006).

may contain links to content that needs to be generated on the fly to appear for the user. Other data loss that could occur may be loss of graphics and the template design.

A snapshot of a Web site usually involves creating a full and accurate copy of an organization's Web site at a particular point in time. A snapshot only provides a picture of a Web site at a particular point in time. A snapshot should include all aspects of the Web site to ensure that a fully functional site can be recreated. The snapshot should include scripts, programs, plug-ins, and browser software—components that make the snapshot fully functional.

A standard Web crawl could be conducted using an open source Web crawler such as Heritrix developed by the Internet Archive for public use. The Heritrix crawler has a long history of support and is designed to respect the robots.txt exclusion directives[42] and META robots tags,[43] and collect material at a measured, adaptive pace unlikely to disrupt normal Web site activity. The advantages of an open source crawler for Web site archiving are that it is non-proprietary and therefore no financial penalties would be incurred. An automated Web crawl could collect data as frequently as the institution desires; initially the crawler could be set to crawl the entire site, and subsequent crawls could collect data from pages that have only been updated since the previous crawl.

To preserve an impression of the Web site at a given moment in time, the institution need only crawl a Web site once or twice a year. This frequency, however would obviously not capture every change made to a Web site, and may miss some of the documented activity that is present. The Web crawler would be implemented to perform infrequent crawls of the Web site. Copies or "snapshots" of the Web site as a whole are taken (ensuring that the functionality of internal links are not destroyed and are maintained). In the meantime, to ensure that the necessary evidence is captured a log of changes that determines when and how documents or Web pages are removed, replaced or updated, is kept. If, for the purposes of accountability and site maintainability, it is important that records of Web site content and changes are made and kept, then this is a viable, inexpensive option.[44] Once again, metadata is the key to effectively managing all records, including records of Web-based activity. (See previous Metadata heading).

[42] For more information on the robots.txt exclusion directives, please visit: www.robotstxt.org/orig.html.
[43] For more information on META robots tags, please visit: www.robotstxt.org/meta.html.
[44] The Web crawl with a log option was researched using "Archiving Web Resources: Guidelines for Keeping Records of Web-based Activity in the Commonwealth Government" from the National Archives of Australia. It is a government recordkeeping document published in March 2001 and can be downloaded from www.naa.gov.au/Images/archWeb_guide_tcm2-903.pdf (last accessed April 28, 2009).

One option for outsourcing the remote harvesting data capture method is presented by the Internet Archive. The Archive-It project is run by the Internet Archive. It is a service provided to smaller organizations that wish to preserve minimal Web content, either from single Web sites or a variety of Web sites. Archive-It partners with the institution and provides a Web-based application that allows users to create, manage and preserve collections of born digital content. Archive—it is run on a subscription basis. The costs associated with the outsourcing option may be prohibitive in terms of financial resources. Subscription rates range from $12,000.00 to $17,000.00 per year.

A further issue that could become problematic for Canadian collecting institutions is the fact that data is stored by the Internet Archive on servers across the globe, including the USA. This means that any data stored is subject to the USA Patriot Act (***U****niting and* ***S****trengthening* ***A****merica by* ***P****roviding* ***A****ppropriate* ***T****ools* ***R****equired to* ***I****ntercept and* ***O****bstruct* ***T****errorism Act*, 2001).[45] Concerns from Canadian Institutions regarding the USA Patriot Act revolve around perceived threats to Canadians' privacy.[46]

An option that copies the Web site, but will not capture associated metadata needed to effectively preserve the digital content of the Web site, is Web site mirroring. A mirror is an exact copy of a data set. It essentially works as a digital "print out" of the Web site. Mirroring of sites occur for a variety of reasons, one of them being to preserve a Web site or Web page.

Mirroring, as stated above, does not capture metadata associated with each Web page file. It is a good option if all the Archives wishes to preserve is evidence of the Web site in question. We offer this solution with the proviso that as there is no metadata capture during the process of mirroring the Web site, there is nothing in place to address evidence of actual records that may appear on the site. We cannot, therefore, recommend Web site mirroring if the collecting archives wishes to preserve evidence of records appearing on the Web site.

Three mirroring tools were researched. The open source crawler HTTrack and a proprietary software program "Grab-a-Site." Both have been utilized effectively in other archival institutions.[47] A further tool was researched that has not been discussed as being successfully implemented by a small or medium sized organization. It is the Adobe Web Capture tool.

[45] USA Patriot Act, 2001. Available at: www.gpo.gov/fdsys/pkg/PLAW-107publ56/pdf/PLAW-107publ56.pdf.
[46] See: CBC News Report on Canada's Privacy Commissioner, Jennifer Stoddart's Annual Report: Patriot Act Seen as Threat to Canadians' Privacy. Available at: www.cbc.ca/canada/story/2006/06/20/privacy-report.html.
[47] E-mail to the Management & Preservation of Electronic Records Listserv, April 3, 2009, from the Electronic Records Archivist at Kentucky Department for Libraries and Archives.

HTTrack is a free and easy-to-use offline browser utility. It allows a user to download a Web site from the Internet to a local directory, building recursively all directories, copying HTML, images and other files from the server to the local directory. HTTrack arranges the original site's relative link-structure. It allows users to simply open a page of the "mirrored" Web site in their browser and to browse the site from link to link, as if viewing it online.[48] This harvester has been used successfully by archivists seeking to preserve Web content in the Microsoft/Windows environment.

Another tool that mirrors the Web site, is the proprietary software "Grab-a-Site" from a US company called Blue Squirrel.[49] The Grab-a-Site software allows the user to download an entire Web site to a hard drive while retaining the original file names and directory structure. Features of the software include its support of many file types (MOV, AVI, JPG, PDF, EXE and ZIP); the ability to export data to enable users to burn data to removable media; the ability to view the site in an easy to navigate view similar to the Windows File Explorer; and it performs relative link adjustments so that if the Web site data is moved the links will still work in subsequent environments.

> The Grab-a-site product information page also stresses the software's capabilities in terms of dynamic Web sites, stating it "grabs sites written in PHP, ASP, JS or Cold Fusion and turns them into static HTML for distribution on Web servers or CD."[50]

The Adobe Web capture tool converts Web pages to PDF files to create PDF versions of the Web page. It is simple to use and therefore easily teachable to staff. It is possible to capture an entire site using Web Capture. Not only do all the links continue to work in the PDF, they also link to local content within the PDF, where applicable, so that you can truly browse the site offline. Web Capture can be invoked through the Acrobat toolbar in Internet Explorer on Windows and through the Adobe Acrobat 9 application on Windows and Mac platforms.

The tool is easy to use, captures various levels of links within a site, has a date and time stamp for captured web pages, and backwards compatibility is assured by Adobe. There is, however, no metadata captured, it reproduces a flat pdf document, which means that it is not possible to remove a portion of page to print, for example a picture, so it becomes necessary to

[48] See the HTTrack Web site for more information: www.httrack.com/.
[49] See the Blue Squirrel Web site for the Grab-a-Site product page: www.bluesquirrel.com/products/grabasite/.
[50] Ibid.

print the whole page, the entire Web site is captured each time, and the tool converts the captured Web site to PDF rather than to PDF/A.

Adobe released the Web capture tool in 2008, we have not heard of successful implementations in similar organizations. It is, however, an extremely simple solution to implement and use. Adobe has a good reputation and a history of support for the client. Adobe tries to ensure that each new product release is backward compatible to several previous versions.

One further tool for Web site archiving (for the archiving of Web sites built with a back-end database to house the information that is generated on-the-fly to the user) is database archiving. The technique is in its infancy, but it is worth describing in some detail as it is a tool that can be utilized to mitigate problems associated with archiving dynamic Web sites using static Web site methods.

Adrian Brown describes the process of archiving database driven Web sites as having three stages:

> first the repository defines a standard data model and format for archived databases; then each source database is converted to the standard format; and, finally a standard access interface is provided to the archived databases.[51]

The Swiss Federal Archives have developed an XML based format that permits long-term preservation of relational database content. The format has a long history of development dating back to the early 1990s. In May 2008 it was accepted as the official format of the European PLANETS project for archiving relational databases. The format is known as SIARD or the Software Independent Archiving of Relational Databases. It preserves data content and metadata as well as the relations in a format that conforms to ISO standards.[52] A briefing paper published in October, 2008 by digital preservation Europe, "Database Preservation: The International Challenge and the Swiss Solution" describes the SIARD process:

[51] Brown, Adrian, *Archiving Web sites* (London: Facet Publishing, 2006), 59.

[52] According to the briefing paper published in October, 2008 by Digital Preservation Europe, "Database Preservation: The International Challenge and the Swiss Solution" (www.digitalpreservationeurope.eu/publications/briefs/database_preservation.pdf), "The use of widely accepted ISO standards ensures to a large extent that stored data could be accessed in the future. Based upon this assumption SIARD records both primary data and metadata automatically in ISO norm formats: SQL1999 UNICODE and most important of them all: XML 1.0. To ensure standardization SIARD converts all proprietary database charters into the equivalent UNICODE character set. Furthermore, SIARD does not archive synonyms as they are not part of the standardized SQL:1999. Sticking to the standards is an iron rule."

> A SIARD archive is a structured non-compressed ZIP container (ZIP-64 standard), permitting practically any file size. It contains two folders: "header" and "content." The header folder stores the database context, the metadata. A single file, *metadata.xml*, assures that we can understand the technical as well as the contextual background of the database. In technical terms SIARD registers on the upmost level (the database) the identifier, the format version, the message digest code of the archiving pc terminal (verifying primary data integrity) etc. On the schema level SIARD stores lists of tables, views and routines. On the table level, SIARD records the constraints and triggers. And as we go deeper into the column level SIARD also specifies the SQL type in use, LOBs (Large Objects) names, and most important of all: foreign keys and candidate keys with referential data – i.e. the relations. At the same time SIARD contextualizes the data. On the database level it lets us register or add (with the SIARD Suite) information on the archive provenance, description, user etc. In lower levels it lets us keep details of the tables and columns names and content. This descriptive information renders the database comprehensible for future users in both contextual and technical terms.
>
> The second folder, content, stores the primary data. The data is archived according to the database structure. For each schema SIARD automatically generates a folder (schema 1, schema 2, etc.), containing the corresponding table series as subfolders (table 1, table 2, etc.). Data itself is stored in XML files (e.g. table1.xml). This schema definition reflects the table's SQL schema metadata. And it specifies that the table is stored as a chain of lines encompassing a sequence of column entries with different XML types. BLOBs and CLOBs (Binary or Character Large Objects containing all sorts of information) are also archived. They are stored in automatically generated folders (e.g. lob1, lob2, etc.) either as TXT or BIN files (record1.text, or record1.bin, etc).[53]

SIARD is also an open format, which would mean that the collecting organization could in fact archive the database without the possible additional costs of obtaining a license to the proprietary content management system required if the Direct Transfer method of capture is employed.

It is clear from the description of SIARD above, that the collecting institution will need to have input from a technologically minded individual to successfully implement the SIARD Suite.

[53] For a more comprehensive discussion of the SIARD format, please see "SIARD Format Description," available for download at the Swiss Federal Archives Web site: www.bar.admin.ch/themen/00532/00536/index.html?lang=en.

It is uncertain at this time if the SIARD Suite is currently available for public use. At a presentation given by Jean-Marc Comment, a representative of the Swiss Federal Archives, to the 16th International Congress on Archives in July of 2008[54] it was noted that the SIARD tools will be available in the future from the Swiss Federal Archives. As of October 12, 2009 nothing appears on the Swiss Federal Archives' Web site[55] regarding the SIARD tools. Due to the uncertainty of availability the SIARD Suite has not been included as a preservation option in this report. It is, however, an interesting option that may be pursued once availability is assured.

Maintaining Web-based Records over Time

Ensuring accessibility to Web-based materials over time raises the same accessibility issues as surround other electronic records. There are steps that can be taken to mitigate these issues including ensuring materials are carefully managed, planning for obsolescence, the use of widely supported standards, implementing security measures to protect against either deliberate or accidental alteration, and ensuring environmental control and monitoring. Most of these steps have been discussed in previous parts of this document, but it is prudent to stress again the importance of these issues.

Careful Management: This might include: maintaining preservation masters and storing these in a separate location; implementing the use of XHTML and avoiding the use of non-standard HTML tags, refreshing storage media regularly, spot checking data to ensure accessibility.

Planning for Obsolescence: Plan for obsolescence by ensuring that records can be copied, reformatted or migrated. Any preservation activities such as the above should be documented in the recordkeeping metadata, including any loss of functionality, content or appearance.

Use of Standards: The importance of the use of standards has been documented above.

Security Measures as a means to Protect data: It is important to build into the Web archiving process security measures that protect data from either deliberate or accidental alteration. Measures can be as simple as keeping the archived data in a secure environment that has controlled access to allow only authorized persons access to the data and providing read-only access to the archived data.

[54] To view the full presentation, please visit: www.planets-project.eu/docs/presentations/ICA2008_Comment_SIARD.pdf.
[55] Swiss Federal Archives Web site: www.bar.admin.ch/index.html?lang=en.

Environmental Control and Monitoring: Best practice dictates that stored media should be kept in optimal temperature and humidity levels, media should be protected against magnetic fields, the use of air filtration units to protect against air pollutants, prohibiting the consumption of food in the storage area, and planning for disasters.

General Action Plan for Web site Preservation

Although there is no generic solution for Web site preservation plans, there are certain elements that will be universal to all programs. When the most appropriate strategy has been identified, a team comprising recordkeeping practitioners, Web site administrators, communications managers, and information technology staff should be selected. The team can develop an overall action plan that includes policies and procedures that is suitable for their needs.

This is a general action plan for Web site preservation that can be adapted to many different institution's needs.

1) Identify recordkeeping requirements for Web-based activity.
2) Determine if existing system satisfies the above requirements or whether it is necessary to design and implement a new system or improve the current system.
3) Raise profile and general awareness within the organization of the general recordkeeping responsibilities of all staff.
4) Carry out risk assessment to determine level of acceptable risk posed.
5) Develop overarching Web site Preservation Policy (or Digital Records Preservation Policy that includes Web site Preservation)[56]
    a) Develop Collection Policy (includes Selection policy)[57]
    b) Develop Selection Policy
        i) Definition of context
        ii) Selection methods
        iii) Selection criteria
            (1) Appraisal of content
            (2) Extent[58]
        iv) Collection list
            (1) Selection of Web resources to collect

[56] The European Electronic Resource Preservation and Access Network (ERPANET) has a useful Digital Preservation Policy Tool. Available at: www.erpanet.org/guidance/docs/ERPANETPolicyTool.pdf. The tool walks institutions through various sections of policy development for digital resources. It defines the benefits of creating a Digital Preservation Policy, scope and objectives, requirements, roles and responsibilities, context, areas of coverage, costs, monitoring and review, and implementation. It also contains a bibliography that points the reader to other well developed Digital Preservation Policies that one can adapt or compare when developing an institutional policy for digital preservation.

[57] Both the policy and collection lists should be reviewed periodically to add or subtract resources.

[58] It is necessary to establish criteria for determining the extent of selected resources – i.e. whether or not external links will be collected.

v) Define boundary definitions
   (1) Determine URL or domain name[59]
   (2) Parameters[60]
vi) Define collection method
vii) Determine timing and frequency of collection – including risk assessment methodology[61]
   (1) Influenced by life-cycle of Web resource
   (2) Rate of content change
   (3) Topicality and significance
viii) Define storage for digital assets[62]

6) Implement Policy
7) Document procedures and processes to ensure strategies are carried out.
8) Begin Web site Preservation Program.
9) Perform checks on captured and stored data.[63]
10) Revisit policy and appraisal objectives on a frequent basis.

Can the action plan chosen for a given body of records be valid for another body of records of the same type, produced and preserved by the same kind of organization, person, or community in the same country?

Yes, any action plan devised for Web site preservation can be easily translated to another similar Web site preservation program in a similar kind of organization. The plan transcends organization, community and nation.

[59] If a single Web resource, such as a page or a document, is to be collected in isolation, then the collection list would simply need to specify the url of that resource. If an entire Web site has been selected this will usually be defined as a domain name.
[60] Parameters define the number of levels of the directory structure to be collected, and whether or not external links should be followed and if so to what depth.
[61] Cornell University has developed a methodology for assessing and mitigating risks to live Web resources: the Cornell Virtual Remote Control Project, available at: http://irisresearch.library.cornell.edu/VRC/. Abstract: The VRC risk management methodology follows a six step process that starts with the identification and the evaluation of Web sites, facilitates the assessment of a site's risk level and strategy building, and initiates a subsequent response. The VRC catalogue seeks to automate this process as much as possible but allows for human control. The stability of a Web site is measured at various risk levels that can be deduced from monitoring a Web site over time regarding its implementation (e.g., HTML tidiness) and its hyperlink structure, as well as from metadata about the Web server (e.g., server software, response time). If a Web site is at high risk it may be necessary to get in touch with the respective site owner. VRC therefore plans to establish recommendations in 'Web content preservability guidelines.' As a last resort the Web resource at risk may also be harvested and preserved to avoid its loss. Abstract from ERPANET: erpaAssessment, available at www.erpanet.org/assessments/show.php?id=1092153049&t=1.
[62] To ensure long-term accessibility of data, it is essential that storage media is refreshed on a regular basis; the action of refreshing storage media should be built into the overall electronic records policy seen in the action plan steps.
[63] Once the Web site has been captured and transferred to the archives environment, checks must be conducted to ensure that all the parts of the Web site captured are working as they should. Checks include, but are not limited to: manually going through and clicking on all hyperlinks; randomly clicking on links; or employing the use of a link testing application to help automate the checking process. Examples of link testing applications are: Link Checker Pro: www.link-checker-pro.com/ Site Audit: www.blossom.com/site_audit.html Cyber Spyder Link Test: www.cyberspyder.com/cslnkts1.html and Link Sleuth: http://home.snafu.de/tilman/xenulink.html.

F. Bibliography

Adobe Web capture tool. Information on the product available at: www.adobe.com/products/acrobat/

Alma Mater Society, "Code of Procedures." Available at: www.amsubc.ca/uploads/government/EXECUTIVE_PROCEDURES_MANUAL_Nov_04.pdf

Alma Mater Society Archives, "Test-Bed Presentation by the AMS Archives to the InterPARES 3 Project Meeting, November 27, 2006." Available at: www.interpares.org/ip3/display_file.cfm?doc=ip3_canada_ubc_ams_research_proposal.pdf

Blue Squirrel, Grab-a-Site Product Page. Available at: www.bluesquirrel.com/products/grabasite

British Columbia (1996), University Act. [RSBC 1996] CHAPTER 468. Available at: www.bclaws.ca/Recon/document/freeside/--%20U%20--/University%20Act%20%20RSBC%201996%20%20c.%20468/00_96468_01.xml

British Columbia, Office of the Information & Privacy Commissioner (Date Assented: 2003), Personal Information Protection Act Regulations. B.C. Reg. 473/2003, O.C. 1234/2003. Available at: www.qp.gov.bc.ca/statreg/reg/P/473_2003.htm

British Columbia, *Society Act*, 4(1)(d). Available at: www.qp.gov.bc.ca/statreg/stat/S/96433_01.htm

Brown, Adrian, *Archiving Web sites* (London: Facet Publishing, 2006).

Brown, Adrian (August 2008), "Digital Preservation Guidance Note 2: Selecting Storage Media for Digital Preservation." Available at: www.nationalarchives.gov.uk/documents/selecting-storage-media.pdf

Brown, Adrian, "Selecting File Formats." Available at: www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf

Canadian Conservation Institute, *Electronic Media Collections Care for Small Museums and Archives*. Available at: www.cci-icc.gc.ca/headlines/elecmediacare/index_e.aspx

Canadian Conservation Institute, *Electronic Media Collections Care for Small Museums and Archives*. Available at: www.cci-icc.gc.ca/headlines/elecmediacare/index_e.aspx

Digital Preservation Europe (October 2008), "Database Preservation: The International Challenge and the Swiss Solution." Available at: www.digitalpreservationeurope.eu/publications/briefs/database_preservation.pdf

European Electronic Resource Preservation and Access Network (ERPANET), *Digital Preservation Policy Tool*, 2003. Available at: www.erpanet.org/guidance/docs/ERPANETPolicyTool.pdf

Greenwood, David J. and Morten Levin, "Reconstructing the Relationships between Universities and Society through Action Research," in Norman K. Denzin and Yvonna S. Lincoln, eds., *The Landscape of Qualitative Research: Theories and Issues*, 2nd ed. (Thousand Oaks: SAGE Publications, 2003), 131-166.

Harvey, Ross, *Preserving Digital Materials* (Munich, Germany: K. G. Saur, 2005).

HTTrack Web site. Available at: www.httrack.com/

Internet Archive, Heritrix Web site. Available at: http://crawler.archive.org

InterPARES 3 Project, TEAM Canada, "Case Study 09 Contextual Analysis: UBC Alma Mater Society," (January 2008). [restricted access document]

InterPARES 3 Project, TEAM Canada, "Case Study 09 – UBC Alma Mater Society: Records Research Questions," (April 2008). [restricted access document]

InterPARES 3 Project, TEAM Canada, "Case Study 09 Alma Mater Society of the University of British Columbia. Workshop 02: Action Item 23: Procedures for Updating Web site content," (November 2008) Available at: www.interpares.org/ip3/display_file.cfm?doc=ip3_canada_cs09_action_23_wks02_v1-2.pdf

InterPARES 3 Project, TEAM Canada, "Case Study 09 Alma Mater Society of the University of British Columbia. Workshop 02: Action Item 24: Web site Content Appraisal," (November 2008). Available at: www.interpares.org/ip3/display_file.cfm?doc=ip3_canada_cs09_wks02_action_24_v1-3.pdf

InterPARES 3 Project, TEAM Canada, "Case Study 09 Alma Mater Society of the University of British Columbia. Workshop 02: Action Item 25: Preservation Process / Strategy," (November 2008). Available at: www.interpares.org/ip3/display_file.cfm?doc=ip3_canada_cs09_wks02_action_25_v1-3.pdf

InterPARES 3 Project, TEAM Canada, "Case Study 09 – Alma Mater Society of the University of British Columbia – Workshop 03 Action Item 21 – Reappraisal of AMS Web site Content," (May 2009). Available at: www.interpares.org/ip3/display_file.cfm?doc=ip3_canada_cs09_wks03_action_21_v1-3.pdf

InterPARES 3 Project, TEAM Canada, "Case Study 09 – Alma Mater Society of the University of British Columbia – Workshop 03 Action Item 22: Technological Option(s)," (May 2009). Available at: www.interpares.org/ip3/display_file.cfm?doc=ip3_canada_cs09_wks03_action_22_v1-2.pdf

InterPARES 3 Project, TEAM Canada, "Case Study 09 – Alma Mater Society of the University of British Columbia – Workshop 03 Action Item 23: On-going Costs of Implementing Identified Technological Options," (May 2009). Available at: www.interpares.org/ip3/display_file.cfm?doc=ip3_canada_cs09_wks03_action_23_v1-3.pdf

InterPARES 3 Project, TEAM Canada, "TEAM Canada Plenary Workshop #03: Action Items and Decisions." [restricted access document]

InterPARES 3 Project, TEAM Canada, "Case Study 09: Alma Mater Society of the University of British Columbia. Technological and Pricing Options for Web site Storage and Capture Summary," (June 2009). Available at: www.interpares.org/ip3/display_file.cfm?doc=ip3_canada_cs09_technological_options_summary_v1-2.pdf

ISO 19005-1:2005 "Document Management - Electronic document file format for long term preservation - Part 1: Use of PDF 1.4 (PDF/A-1)."

Jones, Maggie and Neil Beagrie, *Preservation Management of Digital Materials A Handbook* (London, UK: The British Library, 2001).

Kenney, Anne R., Nancy Y. McGovern, Peter Botticelli, Richard Entlich, Carl Lagoze and Sandra Payette (2002) "Preservation Risk Management for Web Resources. Virtual Remote Control in Cornell's Project Prism," *D-Lib Magazine* 8(1). Available at: www.dlib.org/dlib/january02/kenney/01kenney.html

Lazinger, Susan S., *Digital Preservation and Metadata. History, Theory, Practice* (Englewood, CO: Libraries Unlimited, 2001).

Library of Congress (United States). www.digitalpreservation.gov/

Linden, Jim, Sean Martin, Richard Masters and Roderic Parker, "The Large-Scale Archival Storage of Digital Objects," *DPC Technology Watch Series Report 04-04*, Digital Preservation Coalition (February 2005). Available at: www.dpconline.org/docs/dpctw04-03.pdf

McGovern, Nancy, Anne R. Kenney, Richard Entlich, William R. Kehoe and Ellie Buckley (2004), "Virtual Remote Control. Building a Preservation Risk Management Toolbox for Web Resources," *D-Lib Magazine* 10(4). Available at: www.dlib.org/dlib/april04/mcgovern/04mcgovern.html

McLellan, Evelyn Peters, "General Study 11 Final Report: Selecting Digital File Formats for Long-Term Preservation," InterPARES 2 Project (March 2007). Available at: www.interpares.org/ip2/display_file.cfm?doc=ip2_gs11_final_report_english.pdf

National Archives of Australia, "Archiving Web Resources: Guidelines for Keeping Records of Web-based Activity in the Commonwealth Government," (March 2001). Available at: www.naa.gov.au/Images/archWeb_guide_tcm2-903.pdf

National Archives of Australia. (2004). "Digital recordkeeping self-assessment checklist." Available at: www.naa.gov.au/images/digitalrecordkeepingchecklist_tcm2-923.pdf

Pantalony, Rina Elster, *Protecting your Interests: a legal guide to negotiating Web site development and virtual Exhibition Agreements* (Ottawa, Canada: Minister of Public Works and Governments Services Canada, 1999).

Prom, Christopher J. and Ellen D. Swain (2007), "From the College Democrats to the Falling Illini: Identifying, Appraising, and Capturing Student Organization Web sites," *American Archivist* 70: 344-363.

Smiraglia, Richard P., *Metadata A Cataloger's Primer* (New York, NY: The Hawthorn Press, 2005).

Society of American Archivists. Conference / Workshop Calendar 2009. Available at: http://saa.archivists.org/Scripts/4Disapi.dll/4DCGI/events/ConferenceList.html?Action=GetEvents

Swiss Federal Archives, "SIARD Format Description." Available at: www.bar.admin.ch/themen/00532/00536/index.html?lang=en

G. Glossary

Backward compatibility: Compatible with earlier models or versions of the same product. A new version of a program is said to be backward compatible if it can use files and data created with an older version of the same program. A computer is said to be backward compatible if it can run the same software as the previous model of the computer. [This definition not found in the IP3 Dictionary or Glossary – it has been submitted as a candidate term to the IP3 Terminology Committee]

Client side collection methods: The source from which the Web site is collected for preservation. Client-side collection is collected via the Web browser. [Callow/Shaffer based on Adrian Brown]

Content management system: is a computer application used to manage work flow needed to collaboratively create, edit, review, index, search, publish and archive various kinds of digital media and electronic text. [Wikipedia]

Direct transfer: Acquiring a copy of the data directly from the original source. Requires access to host server. Involves copying data from the server and transferring them to the collecting institution. [Adrian Brown, *Archiving Web sites*]

Internet: the global computer network providing a variety of information and communication facilities to its users, and consisting of a loose confederation of interconnected networks which use standardized communication protocols; (also) the information available on this network. [OED]

Migration: The process of transferring data between storage types, formats, or computer systems. [Wikipedia] Performed to combat **technological obsolescence**.

Preservation Copy: The duplicate of an object, resulting from a reproduction process for preservation purposes. [Callow/Shaffer/IP2]

Remote harvesting: The most common Web archiving technique uses Web crawlers to automate the process of collecting Web pages. Web crawlers typically view Web pages in the same manner that users with a browser see the Web, and therefore provide a comparatively simple method of remotely harvesting Web content. [Wikipedia]

Server-side collection methods: The source from which the Web site is collected for preservation. Server-side collection is collected via the Web server. [Callow/Shaffer based on Adrian Brown]

Technological obsolescence: is a situation where a digital resource is no longer readable because the physical media, the reader required to read the media, the hardware, or the software that runs on it, is no longer available. [Wikipedia]

Traditional media: All non-digital media, including analogue and paper. [Callow/Shaffer]

Version control: the management of changes to documents, programs, and other information stored as computer files.

Web crawler: A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner. [Wikipedia]

Web site: A document or a set of linked documents, usually associated with a particular person, organization, or topic, that is held on such a computer system and can be accessed as part of the World Wide Web. [This definition not found in the IP3 Dictionary or Glossary – it has been submitted as a candidate term to the IP3 terminology committee]

Web site mirroring: A mirror is an exact copy of a data set. It essentially works as a digital "print out" of the Web site. The process of Web site mirroring produces a copy of the original Web site, but does not capture associated metadata. [Callow/Shaffer]

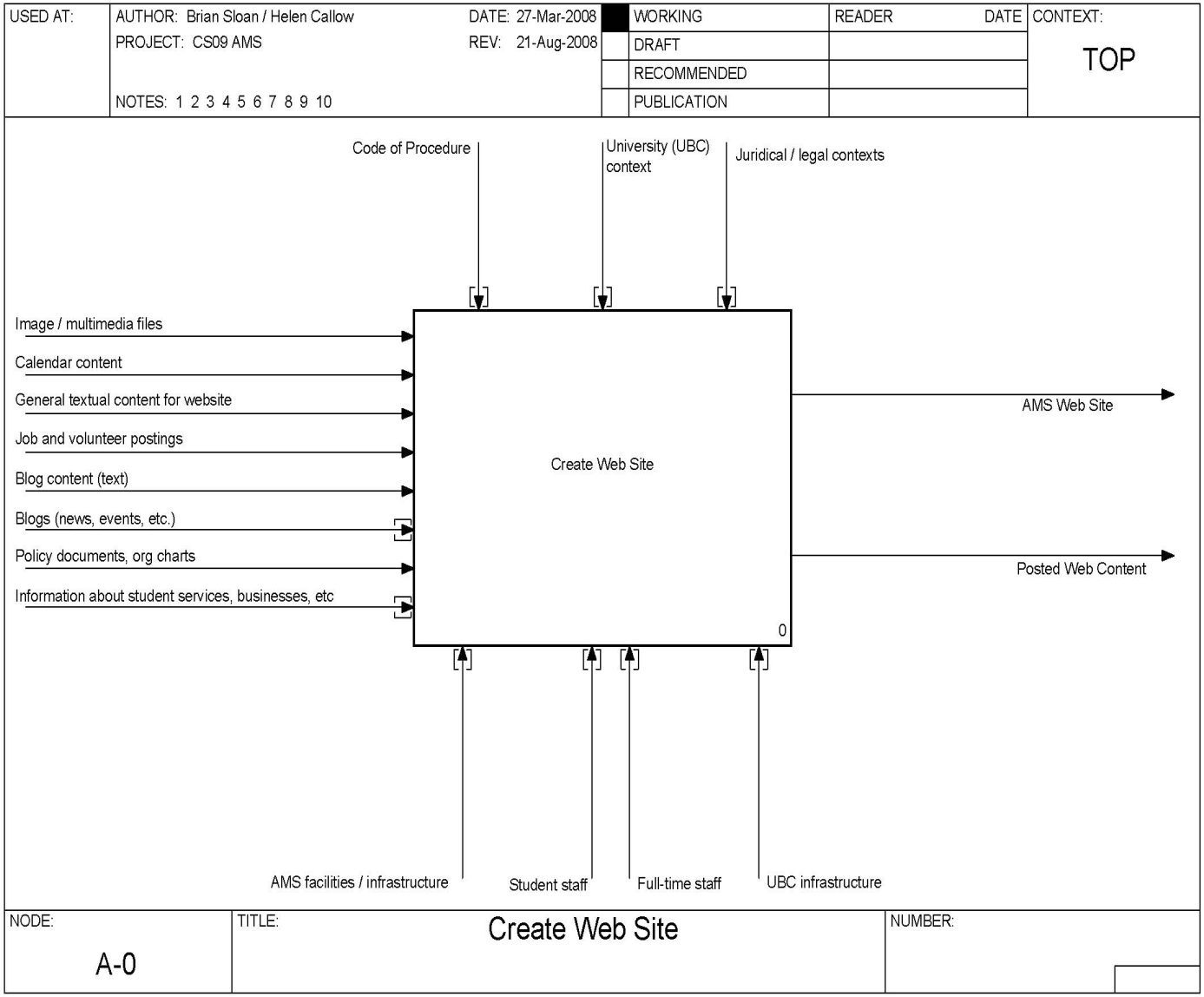
World Wide Web: A system for accessing and retrieving multimedia information over the Internet, whereby documents stored at numerous locations worldwide are cross-referenced using hypertext links, which allow the user to move from one document to another. Also: the network of interlinked information that is accessible via this system. [OED]

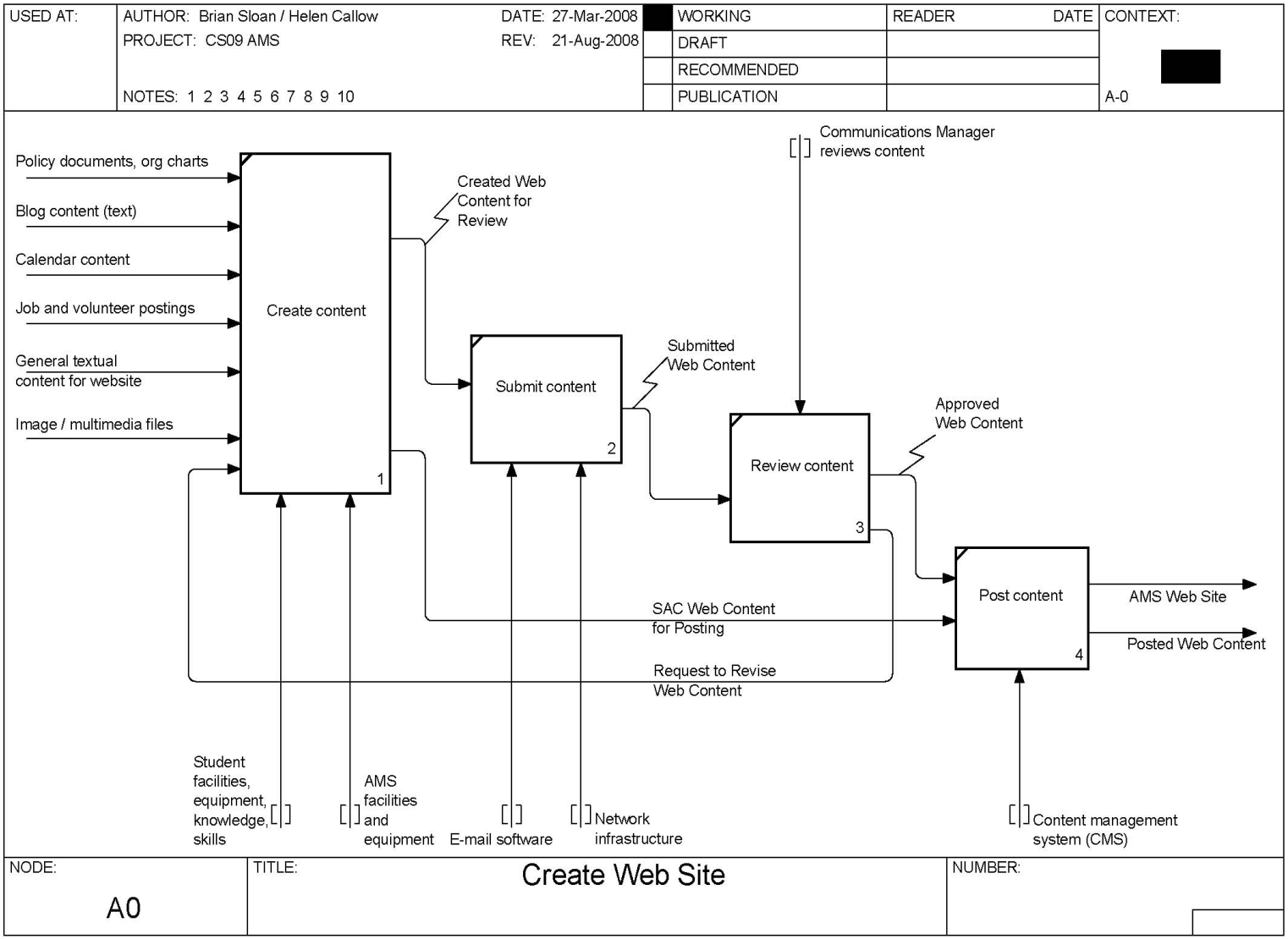
H. IDEF0 model

Case Study 09 (UBC Alma Mater Society) Activity Definitions (v2.0)			
Activity Name	Activity Number	Activity Definition	Activity Note
Create Web Site	A0		
Create content	A1		
Submit content	A2		
Review content	A3		
Post content	A4		

Case Study 09 (UBC Alma Mater Society) Arrow Definitions (v2.0)		
Arrow Name	Arrow Definition	Arrow Note
AMS facilities / infrastructure		
AMS facilities and equipment		
AMS Web Site	(network of linked Web pages)	
Approved Web Content	Web content that has been reviewed and approved and is ready for posting on the public Web site.	
Blog content (text)		
Blogs (news, events, etc.)		
Calendar content		
Code of Procedure		
Communications Manager reviews content		
Content management system (CMS)		
Created Web Content for Review		
E-mail software		
Full-time staff		
General textual content for Web site		

Case Study 09 (UBC Alma Mater Society) Arrow Definitions (v2.0)		
Arrow Name	Arrow Definition	Arrow Note
Image / multimedia files		
Information about student services, businesses, etc.		
Job and volunteer postings		
Juridical / legal contexts		
Network infrastructure		
Policy documents, org charts		
Posted Web Content	Web content that is organized, consolidated and communicated to users via the public Web site.	Content is posted by copying and pasting source files into the Content Management System (CMS).
Request to Revise Web Content	Request sent to Web content creator to revise and resubmit the content.	
SAC Web Content for Posting	Web content created by the Student Administrative Commission, which can be posted directly to the public Web site without any formal review.	
Student facilities, equipment, knowledge, skills		
Student staff		
Submitted Web Content		
UBC infrastructure		
University (UBC) context		


USED AT:
AUTHOR: Brian Sloan / Helen Callow
DATE: 27-Mar-2008
PROJECT: CS09 AMS
REV: 21-Aug-2008
NOTES: 1 2 3 4 5 6 7 8 9 10
WORKING
DRAFT
RECOMMENDED
PUBLICATION
READER
DATE
CONTEXT:
TOP
Code of Procedure
University (UBC) context
Juridical / legal contexts
Image / multimedia files
Calendar content
General textual content for website
Job and volunteer postings
Blog content (text)
Blogs (news, events, etc.)
Policy documents, org charts
Information about student services, businesses, etc
Create Web Site
0
AMS Web Site
Posted Web Content
AMS facilities / infrastructure
Student staff
Full-time staff
UBC infrastructure
NODE:
A-0
TITLE:
Create Web Site
NUMBER:


USED AT:
AUTHOR: Brian Sloan / Helen Callow
DATE: 27-Mar-2008
PROJECT: CS09 AMS
REV: 21-Aug-2008
NOTES: 1 2 3 4 5 6 7 8 9 10
WORKING
DRAFT
RECOMMENDED
PUBLICATION
READER
DATE
CONTEXT:
A-0
Policy documents, org charts
Blog content (text)
Calendar content
Job and volunteer postings
General textual content for website
Image / multimedia files
Create content
1
Created Web Content for Review
Submit content
2
Submitted Web Content
Communications Manager reviews content
Review content
3
Approved Web Content
Post content
4
AMS Web Site
Posted Web Content
SAC Web Content for Posting
Request to Revise Web Content
Student facilities, equipment, knowledge, skills
AMS facilities and equipment
E-mail software
Network infrastructure
Content management system (CMS)
NODE:
A0
TITLE:
Create Web Site
NUMBER:

I. Diplomatic analysis of records

Analysis of the AMS Web site demonstrates that the entities on the Web site do not satisfy all the requirements of a record as defined by InterPARES. The Web site as a whole does not possess a fixed form, there are few formal procedural controls surrounding its creation and maintenance, and the Web site does not possess an explicit archival bond between records within the AMS fonds.

Although the Web site itself is not necessarily a "natural by-product" of any specific business activity of the Society, iterations of the Web site have the potential to be records if they are set aside for future reference or action. There is also the potential for the Web site to contain records that are thought to exist elsewhere due to the lack of formal procedural control. The status of the Web site currently could be characterized as a publication, as it is created and maintained primarily to disseminate information.

J. Conclusions

The experience of the AMS Case Study has led to recommendations that seek to establish the best practices of Web site preservation. The elements that were identified as necessary to accomplish this task include: attention to recordkeeping within an organization; the importance of policies and procedures that govern electronic records creation and preservation; and a certain level of technological ability.

Early on in the AMS case study, TEAM Canada decided that, to streamline and enhance the collection and preservation of the AMS Web site, a procedural document should be produced that outlined the Web site maintenance process. The document should establish a code of practice that transcends the ongoing staff turnover. Ultimately, this document was to be voted on by the organization and implemented. The document produced is appended to this report.

In addition to this document that sets forth procedures for Web site maintenance, it is also recommended that the AMS create and vote on a series of procedures that contain criteria to be followed for what can and cannot be uploaded to the AMS Web site. This would establish precedent that governs Web site content as well as making sure that the AMS organization is aware of restrictions that may be placed on content, due in part to the need to adhere to the Personal Information Protection Act legislation (PIPA). To ensure PIPA is followed, strict criteria regarding the treatment of personal information should be included. Implementing such

procedures will ensure that the AMS is aware of what is present on its Web site, simplifying future appraisals.

It is further recommended that the AMS establish a clear and enforceable recordkeeping system that includes both policies and procedures for the scheduling and transfer of the AMS records into the Archives. The presence of an effective recordkeeping system is crucial to an environment such as the AMS because of the high student/staff turnover. Students and staff, as well as student electives, come and go on an annual basis. A system that contains procedures for the transfer of critical documents to a centralized place, such as the AMS Archives, would benefit incoming staff and executives as they could immediately trace core documents to one location. Until the AMS has a clear, enforceable record keeping system in place that includes a records schedule that determines what is transferred to the Archives, the frequency of transfer, and what is only published on the Web site, it is impossible to carry out a comprehensive appraisal

When an enforceable, efficient recordkeeping system is in place at the AMS, it is recommended that the AMS Archives reappraise its preservation collecting program on an annual basis. This recommendation is because the AMS elects a new student council on an annual basis and each new council may bring with it new ideas on how to effectively use the Web site for dissemination of ideas and policy. An annual reappraisal will ensure that each new initiative is captured and set aside for future needs.

An Action Plan was developed that sets forth the steps recommended to the AMS (see Appendix 1). However, the AMS decided to implement a Web site preservation program before this report was finalized. The AMS decided to use the mirroring option and has begun to capture and save material weekly using the Adobe Web capture tool. However, the AMS currently has none of the recommended policies and procedures in place as put forth in this report.

The AMS chose the Adobe tool as it is inexpensive and supported by a reputable company. During the research process, the IT Manager had stated his reluctance to use Open Source products due to the lack of easily available support, so the familiarity of Adobe and the technical support they offer was a large factor in this decision.

The Adobe tool is used to capture the main AMS Web site once a week. This involves capturing a few pages that are beyond the AMS domain as some parts of the AMS Web site that

look to be a part of the whole are actually separate sites from the AMS domain.[64] The tool has also been set up to capture non-AMS Web sites that show AMS activities. The tool is being used to capture the AMS page on Facebook once a week and there are also AMS on Flickr and AMS on Twitter to consider. The tool is also being used for monthly captures of the AMS Constituency Web sites.[65] The Constituency Web sites are quite large, and are subject to infrequent changes, hence the monthly captures. The time line for Constituency sites will be re-evaluated in September.[66]

Once a capturing method had been chosen, a storage solution was implemented. The archivist and IT Manager determined that the main Web site and the AMS on Facebook page requires 70MB of storage space. Saving this data weekly requires 3500MB or 3.5GB per year. The monthly capture of the Constituency Web sites adds 2GB per year to the storage requirements. The AMS Archivist informed the GRAs that a computer in the Archives that is currently unused has 100GB of free storage space, "so we are saving the Web site there, and would have enough room for 30 years."[67] The Archives has also purchased an external hard drive and copy the captured Web site on to that as a backup.

The procedure is as follows:[68]

> Every Tuesday, the Archives Assistant institutes a capture of the AMS main Web site, the AMS pages that are outside of the main Web site domain, but are part of the whole AMS Web site, and the AMS on Facebook pages.
>
> Web site capturing for the main site takes approximately 40 minutes; however, the Web pages outside of the domain and the Facebook page are captured manually. Time spent manually capturing the additional Web pages is minimal because there are so few pages, but this would not be a feasible solution if multiple pages were to be captured in this way.
>
> The Constituency Web sites are captured monthly.

[64] These separate pages are Catering, Conferences, and Elections. The URL's are slightly different from the main AMS Site and consist of static HTML Pages as opposed to the PHP generated pages that make up the main AMS Web site.
[65] At the time of InterPARES' involvement with the AMS as a case study none of these extra Web sites were a concern, the archivist only wanted to capture and preserve the main Web site which is what we focused our research on.
[66] AMS Constituencies are the student associations and undergraduate societies of the degree granting faculties and schools of UBC. Every constituency gets at least one seat on the AMS Council, and one extra seat for every 1,500 students. Constituencies have their own elected councils and can assist students with the day to day concerns and issues specific to their school or faculty. There are 21 Constituency groups, 15 of which have their own Web site. The URL that lists the constituency groups is: www2.ams.ubc.ca/index.php/student_government/category/constituencies
[67] E-mail from Sheldon Goldfarb to Helen Callow and Elizabeth Shaffer, August 4, 2009.
[68] E-mail from Sheldon Goldfarb to Helen Callow, August 6, 2009.

It is the AMS's intent to save the captured data indefinitely.

Ideally, the recommended policies and procedures would be adhered to. There are real dangers involved by acting without such regulations in place, for example PIPA legislation could be violated and more records could be lost without the presence of an effective recordkeeping system. With regard to the actual Web site preservation program, without adhering to the steps for the Web site that we define (file format specifications, file naming specifications and the addition of metadata to uploaded files) the Web site may contain unreadable files that the Adobe tool captures and saves. Without these specifications in place, the only way to ensure that all captures are working is for every captured file to be checked.

The AMS's storage choices may become challenging. A solution that relies solely on hard drive storage is problematic. Joe Iraci of the Canadian Conservation Institute states that, "hard drives are not for long-term storage and data needs to be moved to a new hard drive every 2 to 5 years."[69] The AMS needs to keep a close eye on data stored on the hard drives it has designated for storage to ensure that the media do not become obsolete and/or the data corrupted.

To preserve the captured data indefinitely may not be advisable. The AMS needs to ensure that a comprehensive cataloguing scheme is in place before it is swamped in a quagmire of data, with one week's capture indistinguishable from the next. Pursuing this path is also going to ensure high storage costs as more and more data is captured and needs to be stored and migrated.

It is not clear whether the AMS intends to follow InterPARES advice by re-evaluating its Web site capture time lines on a regular basis. This, as stated previously, is necessary due to the annual turnover of the AMS Executive as each incoming administration may engage the Web site in a different way from previous administrations.

The creation and maintenance of authentic, reliable, accurate and durable evidence of Web-based activity is essential if the AMS is to retain institutional memory and meet community expectations. The organizational culture has been identified as a factor in the effective implementation of any solution. Recommendations are to implement policies and procedures that will define the AMS's recordkeeping activities and in turn, help streamline the Web site preservation program.

[69] E-mail from Joe Iraci to Randy Preston, May 18, 2009.

Appendix 1: AMS Action Plan for Web Site Preservation

1) Devise retention schedules that govern AMS records. Including all data associated with the archiving of the AMS Web site.[70]
2) Devise procedures that govern Web content for upload.[71]
3) Implement procedural document that outlines the Web site maintenance process seen above.[72]
4) Identify recordkeeping requirements for Web-based activity.
5) Determine if existing system satisfies the above requirements or whether it is necessary to design and implement a new system or improve current system.
6) Raise profile and general awareness within the organization of the general recordkeeping responsibilities of all staff.
7) Carry out a risk assessment to determine level of acceptable risk posed.
8) Develop overarching Web site Preservation Policy (or Digital Records Preservation Policy that includes Web site Preservation).[73]
    a) Develop Collection Policy (includes Selection policy)[74]
    b) Develop Selection Policy
        i) Definition of context
        ii) Selection methods
        iii) Selection criteria

[70] Web pages should be subject to the same records management controls as other electronic records, since they provide evidence of the online activities of the AMS. In addition to improved overall recordkeeping, the AMS would benefit in terms of costs associated with storage of electronic media if effective disposition schedules were in place in the organization.
[71] It is recommended that the AMS create and vote on a series of procedures that contain criteria to be followed for what can and cannot be uploaded to the AMS Web site. This would establish precedent that governs Web site content as well as making sure that the AMS organization is aware of restrictions that may be placed on content. Implementing such procedures will ensure that the AMS is aware of what is present on its Web site, simplifying future appraisals. As the AMS is legally bound by the Personal Information Protection Act (PIPA) legislation, strict criteria regarding the treatment of personal information should be included.
[72] Procedures should include: metadata creation, file format specifications, and file naming specifications. Metadata and file format specification has already been discussed within this document. However, it is our recommendation to convert all files for upload to the PDF/A format. PDF/A is a file format that has been developed for the long-term preservation of electronic documents, is an ISO standard and conforms to most of the criteria that Adrian Brown distinguishes as important. It is an accessible format for both MAC and PC users, although proprietary it has a long history of support, allows for backward compatibility, and is the de facto standard of file formats. It should be noted however, that the AMS archives should also retain the original documents posted to the Web site as well as paper copies of extremely important documents. File naming specifications should also be uniformly implemented for those documents uploaded to the AMS Web site. Such uniformity will allow for lessening version control errors as well as ensuring that the documents posted do not posses file names that contain elements that will cause the Web site to break when attempting to read the files, such as capital letters, spaces and punctuation. A suggested naming format is as follows: committee_or_group_name_name_of_document_date. For example: ams_finance_commission_budget_april_2009.
[73] The European Electronic Resource Preservation and Access Network (ERPANET) has a useful Digital Preservation Policy Tool. Available at: www.erpanet.org/guidance/docs/ERPANETPolicyTool.pdf. The tool walks institutions through various sections of policy development for digital resources. It defines the benefits of creating a Digital Preservation Policy, scope and objectives, requirements, roles and responsibilities, context, areas of coverage, costs, monitoring and review, and implementation. It also contains a bibliography that points the reader to other well developed Digital Preservation Policies that one can adapt or compare when developing an institutional policy for digital preservation.
[74] Both the policy and collection lists should be reviewed periodically to add or subtract resources.

(1) Appraisal of content
(2) Extent[75]

iv) Collection list
(1) Selection of Web resources wish to collect

v) Define boundary definitions
(1) Determine URL or domain name[76]
(2) Parameters[77]

vi) Define collection method

vii) Determine timing and frequency of collection – including risk assessment methodology[78]
(1) Influenced by life-cycle of Web resource
(2) Rate of content change
(3) Topicality and significance

9) Define storage for digital assets[79]
10) Implement Policy.
11) Document procedures and processes to ensure strategies are carried out.
12) Begin Web site preservation program.
13) Perform checks on captured and stored data.[80]
14) Revisit policy and appraisal objectives on an annual basis.[81]

[75] It is necessary to establish criteria for determining the extent of selected resources – i.e. whether or not external links will be collected.
[76] If a single Web resource, such as a page or a document is to be collected in isolation, then the collection list would simply need to specify the url of that resource. If an entire Web site has been selected this will usually be defined as a domain name.
[77] Parameters define the number of levels of the directory structure to be collected, and whether or not external links should be followed and if so to what depth.
[78] Cornell University has developed a methodology for assessing and mitigating risks to live Web resources: the Cornell Virtual Remote Control Project available at http://irisresearch.library.cornell.edu/VRC/. Abstract: The VRC risk management methodology follows a six step process that starts with the identification and the evaluation of Web sites, facilitates the assessment of a site's risk level and strategy building, and initiates a subsequent response. The VRC catalogue seeks to automate this process as much as possible but allows for human control. The stability of a Web site is measured at various risk levels that can be deduced from monitoring a Web site over time regarding its implementation (e.g., HTML tidiness) and its hyperlink structure, as well as from metadata about the Web server (e.g., server software, response time). If a Web site is at high risk it may be necessary to get in touch with the respective site owner. VRC therefore plans to establish recommendations in 'Web content preservability guidelines'. As a last resort the Web resource at risk may also be harvested and preserved to avoid its loss. Abstract from ERPANET: erpaAssessment at www.erpanet.org/assessments/show.php?id=1092153049&t=1.
[79] To ensure long-term accessibility of data it is essential that storage media is refreshed on a regular basis, the action of refreshing storage media should be built into the overall electronic records policy seen in the action plan steps. It is also important to highlight the fact that if the AMS stores each iteration of its Web site indefinitely then the costs associated with refreshing the media will soar over time as the data collected grows. This ought to be taken into consideration when devising overall recordkeeping schedules.
[80] Once the Web site has been captured and transferred to the AMS archives environment] checks must be conducted to ensure that all the parts of the Web site captured are working as they should. Checks include, but are not limited to: manually going through and clicking on all hyperlinks; randomly clicking on links; or employing the use of a link testing application to help automate the checking process. Examples of link testing applications are: Link Checker Pro: www.link-checker-pro.com/ Site Audit: www.blossom.com/site_audit.html Cyber Spyder Link Test: www.cyberspyder.com/cslnkts1.html and Link Sleuth: http://home.snafu.de/tilman/xenulink.html.
[81] This recommendation takes into account the fact that the AMS elects a new student council on an annual basis. Each new council may bring with it new ideas on how to effectively use the Web site for dissemination of ideas and policy. An annual reappraisal will ensure that each new initiative is captured and set aside for future research needs and will take into account any mandated changes that are implemented to procedures.

The AMS decided to implement a Web site preservation program before this report was finalized. The AMS has decided to use the mirroring option and has begun to capture and save material weekly using the Adobe Web capture tool. However, the AMS currently has none of the recommended policies and procedures in place as put forth in this report.

The AMS chose the Adobe tool as it is inexpensive and supported by a reputable company. During the research process, the IT Manager had stated his reluctance to use Open Source products due to the lack of easily available support, so the familiarity of Adobe and the technical support they offer was a large factor in this decision.

The Adobe tool is used to capture the main AMS Web site once a week. This involves capturing a few pages that are beyond the AMS domain as some parts of the AMS Web site that look to be a part of the whole are actually separate sites from the AMS domain.[82] The tool has also been set up to capture non-AMS Web sites that show AMS activities. The tool is being used to capture the AMS page on Facebook once a week and there are also AMS on Flickr and AMS on Twitter to consider. The tool is also being used for monthly captures of the AMS Constituency Web sites.[83] The Constituency Web sites are quite large, and are subject to infrequent changes, hence the monthly captures. The time line for Constituency sites will be re-evaluated in September.[84]

Once a capturing method had been chosen, a storage solution was implemented. The archivist and IT Manager determined that the main Web site and the AMS on Facebook page requires 70MB of storage space. Saving this data weekly requires 3500MB or 3.5GB per year. The monthly capture of the Constituency Web sites adds 2GB per year to the storage requirements. The AMS Archivist informed the GRAs that a computer in the Archives that is currently unused has 100GB of free storage space, "so we are saving the Web site there, and would have enough room for 30 years."[85] The Archives has also purchased an external hard drive and copy the captured Web site on to that as a back up.

[82] These separate pages are Catering, Conferences, and Elections. The URL's are slightly different from the main AMS Site and consist of static HTML Pages as opposed to the PHP generated pages that make up the main AMS Web site.
[83] At the time of InterPARES' involvement with the AMS as a case study none of these extra Web sites were a concern, the archivist only wanted to capture and preserve the main Web site which is what we focused our research on.
[84] AMS Constituencies are the student associations and undergraduate societies of the degree granting faculties and schools of UBC. Every constituency gets at least one seat on the AMS Council, and one extra seat for every 1,500 students. Constituencies have their own elected councils and can assist students with the day to day concerns and issues specific to their school or faculty. There are 21 Constituency groups, 15 of which have their own Web site. The URL that lists the constituency groups is: www2.ams.ubc.ca/index.php/student_government/category/constituencies
[85] E-mail from Sheldon Goldfarb to Helen Callow and Elizabeth Shaffer, August 4, 2009.

The procedure is as follows:[86]

> Every Tuesday, the Archives Assistant institutes a capture of the AMS main Web site, the AMS pages that are outside of the main Web site domain, but are part of the whole AMS Web site, and the AMS on Facebook pages.
>
> Web site capturing for the main site takes approximately 40 minutes; however, the Web pages outside of the domain and the Facebook page are captured manually. Time spent manually capturing the additional Web pages is minimal because there are so few pages, but this would not be a feasible solution if multiple pages were to be captured in this way.
>
> The Constituency Web sites are captured monthly.

It is the AMS's intent to save the captured data indefinitely.

Ideally, the recommended policies and procedures would be adhered to. There are real dangers involved by acting without such regulations in place, for example PIPA legislation could be violated and more records could be lost without the presence of an effective recordkeeping system. With regard to the actual Web site preservation program, without adhering to the steps for the Web site that we define (file format specifications, file naming specifications and the addition of metadata to uploaded files) the Web site may contain unreadable files that the Adobe tool captures and saves. Without these specifications in place, the only way to ensure that all captures are working is for every captured file to be checked.

The AMS's storage choices may become challenging. A solution that relies solely on hard drive storage is problematic. Joe Iraci of the Canadian Conservation Institute states that, "hard drives are not for long-term storage and data needs to be moved to a new hard drive every 2 to 5 years."[87] The AMS needs to keep a close eye on data stored on the hard drives it has designated for storage to ensure that the media do not become obsolete and/or the data corrupted.

To preserve the captured data indefinitely may not be advisable. The AMS needs to ensure that a comprehensive cataloguing scheme is in place before it is swamped in a quagmire of data, with one week's capture indistinguishable from the next. Pursuing this path is also going to ensure high storage costs as more and more data is captured and needs to be stored and migrated.

[86] E-mail from Sheldon Goldfarb to Helen Callow, August 6, 2009.
[87] E-mail from Joe Iraci to Randy Preston, May 18, 2009.

It is not clear whether the AMS intends to follow InterPARES advice by re-evaluating its Web site capture time lines on a regular basis. This, as stated previously, is necessary due to the annual turnover of the AMS Executive as each incoming administration may engage the Web site in a different way from previous administrations.

The creation and maintenance of authentic, reliable, accurate and durable evidence of Web-based activity is essential if the AMS is to retain institutional memory and meet community expectations. The organizational culture has been identified as a factor in the effective implementation of any solution. Recommendations are to implement policies and procedures that will define the AMS's recordkeeping activities and in turn, help streamline the Web site preservation program.

Appendix 2: Procedural Document Governing Web Site Creation and Maintenance

Updates and maintenance are performed in an informal, ad-hoc basis on the AMS Web site. Current practice includes verbal requests for changes made in passing to the Web editor, who then performs the changes without review. Other changes come in the form of email requests sent to the Communications manager, who reviews the requested change for grammatical type errors and once satisfied, forwards the request to the Web editor, who again, uploads the changes to the Web site using the Whitematter templates.[1] The process at the AMS holds that the communications manager should vet all changes, but this practice does not always occur. The only section of the Web site in which this process does not transpire is the pages that contain content from the Student Administration Council (SAC) who are empowered to update content on their own, using the same formula as the AMS Web editor.

Best practice suggests that any updates/maintenance to a Web page should go through several review processes before the content is made live for the public to view. The diagram showing the Web update process, addresses these ideals in the formal process.

However, as it is difficult as an outsider to mandate to others procedures for how they work, it is more important for the InterPARES team to address the archival procedures that need to take place at the mid-point in the process, and make sure that these are followed. The diagram showing the Archival process, lays out the steps for this process to occur.

The Web update process has a number of inputs throughout the process. These are explained as follows:

A. Documents and or requests that define a Web site update.
B. Archival Criteria. Internal standards that identify items that need to be added to the Web archive.
C. Archival Development Processes. The document that describes the Web site archiving process.
D. Internal and Customer requirements. Global (generally internal) and update specific (generally customer) standards for Web site updates in general and the specific update.
E. Archival Requirements. Web site archival standards.

The actual processes to be followed are:

1. Web site owner receives an update Candidate.

[1] InterPARES 3 Project, TEAM Canada, "Case Study 09 – UBC Alma Mater Society: Records Research Questions," (April 2008)," question 4.

2. Candidate is reviewed for suitability for inclusion into the Web site.
3. Candidate is placed in the development queue.
4. Development team checks candidates include everything needed for production (i.e. suitable complete content)
5. Build out or convert update to Web format
6. Check update to see if it needs to be included into the Web site archive.
7. Place update on Development (staging area) Web site.
8. Run quality assurance on Development Web site.
9. Move development site changes to Production (public site)
10. Run quality assurance on Production Web site.

Again, currently a staging area is not used and content is added to the live site on an ad-hoc basis. The criterion that needs to be emphasized in order for the arching of the Web site to be achieved is set forth in the Archival Process document.

Inputs for this section are:

A. Archival Criteria. Internal standards that identify items that need to be added to the Web archive.

And the processes for achieving this goal are as follows:

1. Archival Process owners receive update for Web site archive.
2. The update is checked to determine that all archive content requirements are met.
3. Place Metadata in update.[2]
4. The update Metadata is checked to determine that the format meets archive requirements.
5. Return update to Development Queue.

The archival criteria are:

1. Ensure that the settings on the Content Management System, (AFAIR for Expression Engine) are set to allow crawls (If remote harvester is used for capture
2. If direct transfer is used for collection, ensure that the hyperlinks within the archived site are adjusted from absolute links to relative links; and the appropriate search engine (the one used in the original environment) is installed in the new environment to ensure that search functionality is preserved.

[2] Within the metadata concerned with the Archival process, those pages not marked for preservation could be tagged using the robots.text meta tag to exclude the crawler from these pages.

Web Update Process:

Inputs

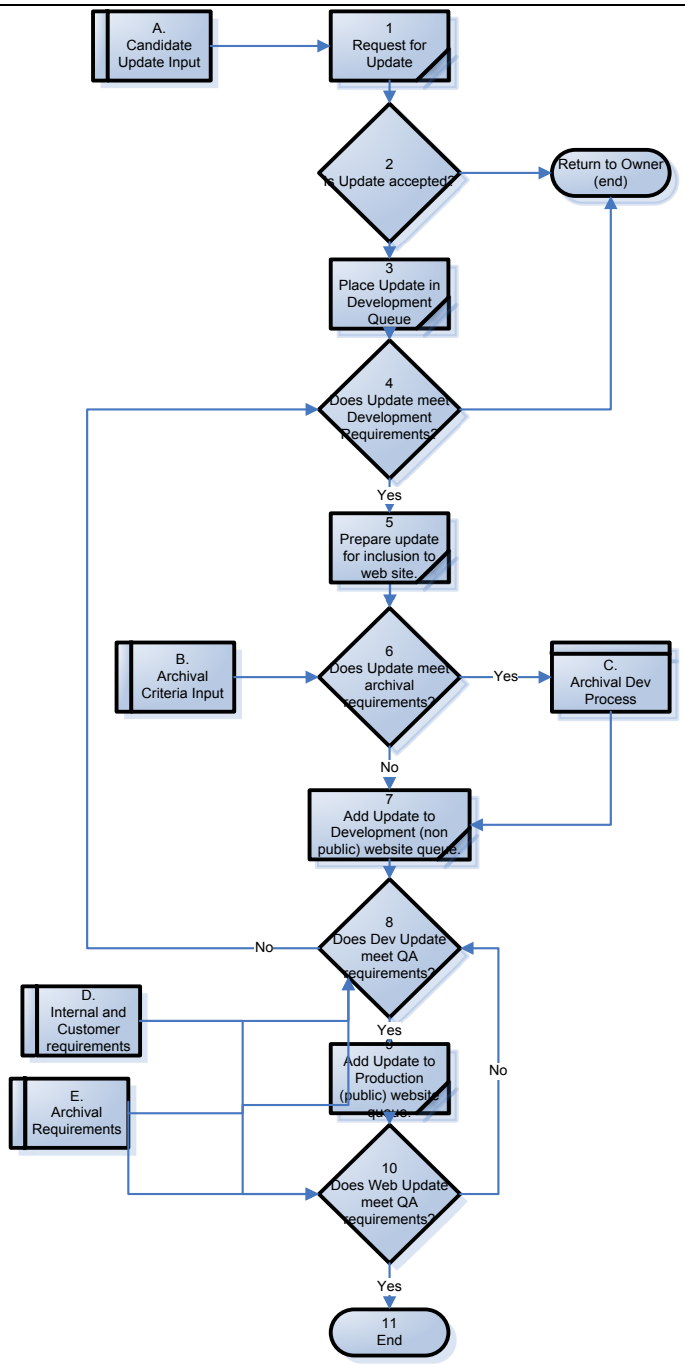
A) Documents and or requests that define a Web site update.

B) Archival Criteria. Internal standards that identify items that need to be added to the Web archive.

C) Archival Development Processes. The document that describes the Web site archiving process.

D) Internal and Customer requirements. Global (generally internal) and update specific (generally customer) standards for Web site updates in general and the specific update.

E) Archival Requirements. Web site archival standards.



Processes

1) Web site owner receives an update Candidate.

2) Candidate is reviewed for suitability for inclusion into the Web site.

3) Candidate is placed in the development queue.

4) Development team checks candidates include everything needed for production (i.e., suitable complete content)

5) Build out or convert update to Web format

6) Check update to see if it needs to be included into the Web site archive.

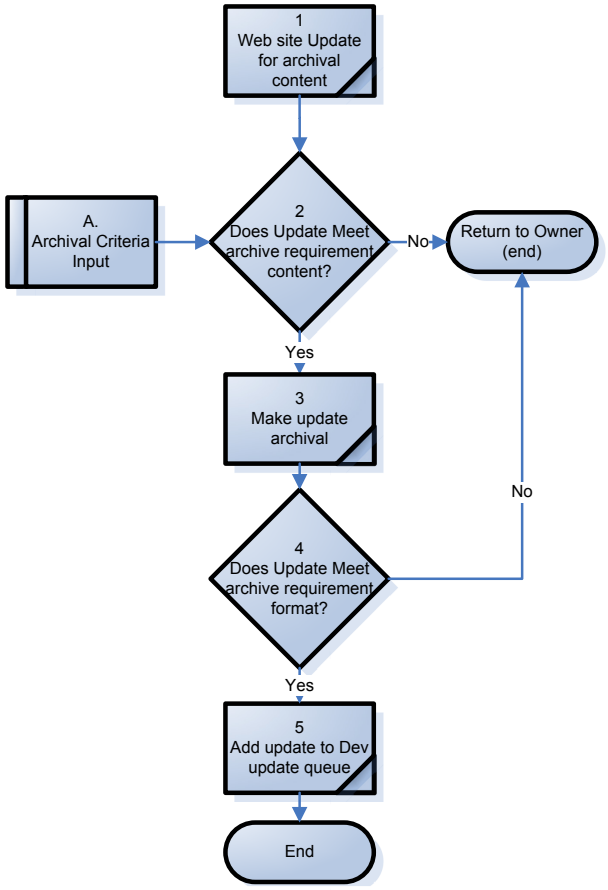
7) Place update on Development (staging area) Web site.

8) Run quality assurance on Development Web site.

9) Move development site changes to Production (public site)

10) Run quality assurance on Production Web site.

Archival Process:

Inputs		Processes
A) Archival Criteria. Internal standards that identify items that need to be added to the Web archive.		1) Archival Process owners receive update for Web site archive. 2) The update is checked to determine that all archive content requirements are met. 3) Place Metadata in update. 4) The update Metadata are checked to determine that the format meets archive requirements. 5) Return update to Development Queue.