InterPARES 3 Project
International Research on Permanent Authentic Records in Electronic Systems
TEAM Canada

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

Technological and Pricing Options for Web site Storage and Capture

Status: Final (public)

Version: 1.2

Date Submitted: June 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

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

Project Unit: Research

URL: http://www.interpares.org/ip3/display_file.cfm?doc=ip3_canada_cs09_technological_options_summary_v1-2.pdf

Document Control

Version history			
Version	Date	By	Version notes
1.0	2009-06-09	H. Callow, E. Shaffer	Combined and condensed version of the Wks03 Action Item 22 and Action Item 23 reports, prepared at the request of S. Goldfarb.
1.1	2009-06-25	R. Preston	Minor content and copy edits.
1.2	2013-05-23	R. Preston	Minor content and copy edits for public version.

Table of Contents

Introduction

The University of British Columbia (UBC) Alma Mater Society (AMS) approached the InterPARES 3 Project (IP3) with a view to preserve information contained within the AMS Web site. They need simple procedures in place so that the archives' assistants can easily be taught to implement and perform any recommended solutions. As a small organization, the AMS has limited resources, so need some easy, cost-effective ways to capture and preserve the Web site.

Previous case study documents (the records research questions) identified the Web site as dynamic (Web pages generated "on the fly" from smaller elements of content). "The Web site uses PHP to pull data out of a MYSQL database and format and present this data "on the fly" to users as navigable HTML Web pages."[1] As this is the case, the best method for collecting data is using server-side collection (the direct transfer method outlined below).

We offer client-side collection methods as alternatives within this report as the AMS Archivist wishes to preserve only an impression of the Web site content and is not particularly concerned at this time with the presence or absence of records contained on the Web site, or evidence of the mandates of Council being adhered to.

Three possible solutions have been identified in order for the AMS to archive its Web site taking into account the AMS environment. Direct transfer of the Web site data from the original hosting source is one solution. Another is remote harvesting of data. 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. The third solution is to produce a mirror of the Web site. All scenarios are explored further below.

Web Site Capture Methods

<u>Direct Transfer Option</u>

Pros: Authentic, reliable record of the Web site is captured for preservation; only solution that addresses the dynamic nature of the Web site

Cons: Highly technical solution; possibly expensive; for ease of use the entire Web site will need to be transferred each time which will lead to higher storage costs and more data to examine when checks are performed or used to fulfill a research request.

The only way to fully recreate the AMS Web site in a preservation environment is through Direct Transfer of data. Direct transfer works by acquiring a copy of the data, in this case the AMS Web site, 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.[2] To ensure that the archived Web site is as authentic as possible, a

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

[2] 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

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.

Currently, the AMS is in the process of migrating its Web site onto the AMS server. This will make a recreation of the technological environment somewhat easier for the AMS as it is already known to the Information Technology Manager who will be responsible for setting up the recreation on to a preservation server. At this time, questions remain regarding the cost of obtaining additional licenses in order to copy the database or content management system into the preservation environment.[3] This cost could impact the viability of such action.

Direct transfer is the only method that takes into consideration the dynamic nature of the 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. This could be a barrier for the AMS in order to implement direct transfer as a capturing method as the Archivist would prefer a less technical solution due to high staff turnover and the need to train each incoming staff member.

Frequency of collection would also determine whether or not this is a viable option in terms of the constraints in place. Effective use of human resources would mean the transfer of all of the files each time data is collected. This however, would result in a need for greater storage space, and the associated costs that go along with this. Due to these costs, it would be virtually impossible for the AMS Archives to keep every instance of the Web site that is collected indefinitely. Retention schedules would have to be devised and implemented that govern the disposition of the Web site instances that are preserved.

Remote Harvesting Options

We identified three options using the remote harvesting method that suit the needs of the AMS. A standard web crawl, a standard web crawl plus the addition of a log that documents Web site changes and an option of outsourcing the web crawl to a third party.

As stated previously, 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. Also, using this method may result in the presence of broken links within the copied data environment as pages 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.

The AMS should consider adding metadata to its Web site content upload processes found in the procedural document created for InterPARES in 2008. Without the addition of metadata to the

search functionality is preserved. For a more comprehensive explanation please see: Adrian Brown, *Archiving Websites* (London: Facet Publishing, 2006).

[3] An additional cost may be incurred with regard to the purchase of an extra license to allow the content management system to be installed within the preservation environment. A communiqué has been sent to Whitematter to clarify any additional costs that may be associated with licensing in the preservation environment. As of this time (June 8, 2009) there has been no response from Whitematter.

uploaded Web content, we cannot in good faith recommend using a remote harvesting option to archive the site completely.

Standard Automated Web crawl

Pros: Once implemented this is an automated process; open source web crawlers are free to use and many (including the one we are recommending Heritrix) have a long history of support and are used by well known institutions such as the Library of Congress; initial capture would be of the entire site, subsequent captures would be only those parts of the site that have changed (less storage requirements);

Cons: Does not address dynamic nature of the Web site; will not capture all possible user generated pages; archived site may contain broken links; limited capture of Web site metadata will not address issues concerning presence of records on the site

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 advantages of an open source crawler for the AMS are that it is non-proprietary and therefore no financial penalties would be incurred.[4] An automated Web crawl could collect data as frequently as the AMS desires; initially the crawler could be set to crawl the entire site, and subsequent crawls could collect data only from pages that have been updated since the previous crawl.

The frequency of the crawl would be determined by what information the AMS wishes to preserve. Workshop 03 Action Item 21 identified time periods that may be important for the AMS to document such as student elections.[5] This option would lower storage costs as the whole site need not be captured each time, however, it would increase the human resource requirements needed to implement as well as the upkeep requirements associated with an open source crawler.[6]

In order to preserve an impression of the Web site at a given moment in time, the AMS need only crawl the site once or twice a year. This frequency, however would obviously not capture every change made to the Web site, and may still miss some of the documented activity that is present on the AMS Web site.

[4] The AMS Information Technology manager stated his reluctance to implement an open source option at a meeting on April 9, 2009. His reasoning is that open source is not user friendly, and provides no support, therefore, he anticipates having to spend an exorbitant amount of time trouble shooting any open source option. Open source is recommended in this report due to the financial constraints expressed by the AMS.

[5] As noted in the "Workshop 03 Action Item 21 – Reappraisal of AMS Web Site Content" report (p. 5), "Content can change on the AMS Web site on an almost daily basis; most of these changes are semantic and therefore are not necessary to save in their daily iterations. However, certain times throughout the year have been identified as being more consequential. For example, the beginning of Term two of the winter semester is an important time within the AMS organization with the student elections taking place. The Web site is increasingly becoming a major communication device in terms of candidates speaking to their voting public. At this time of year the Web site changes quite dramatically as candidate biographies are published as well as events connected to elections and the final electoral results" (http://www.interpares.org/ip3/display_file.cfm?doc=ip3_canada_cs09_wks03_action_21_v1-3.pdf).

[6] However, the AMS Archives needs to take into consideration another finding shown in the Action Item 21 report that advices that the arrival of the newly elected AMS executive brings a commitment to utilize the AMS Web site more frequently. Recent communications with the AMS Communications Manager have discovered that significant content changes to the News and Executive Blog sections of the Web site occur on a weekly basis. The AMS Archives will need to determine which of these communiqués are important to capture and place within their preservation program.

We would recommend using the Heritrix crawler from the Internet Archive. It has a long history of support and is designed to respect the robots.txt exclusion directives[7] and META robots tags,[8] and collect material at a measured, adaptive pace unlikely to disrupt normal Web site activity. The AMS must contact the Web site Content Management Provider to ensure that access is given to the spider by removing robots.txt exclusion directives and META robots tags from metadata.

Standard Automated Web crawl plus log

Pros: inexpensive; history of support; Web crawl automated; less storage requirements; would account for evidence of records on the site

Cons: Does not address dynamic nature of the Web site; will not capture all possible user generated pages; archived site may contain broken links; limited capture of Web site metadata

A standard Web crawl could be implemented to address the mandated actions set forth in Workshop 03 Action Item 21.[9] 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.[10]

The outsourcing solution

Pros: Non-technical solution; easily taught to Archives Assistants; Outsource partner performs crawl; Outsource partner stores data

Cons: Expensive; Outsource partner controls data; Data stored on US servers (PIPA concerns)

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.

[7] For more information on the robots.txt exclusion directives, please visit: http://www.robotstxt.org/orig.html.

[8] For more information on META robots tags, please visit: http://www.robotstxt.org/meta.html.

[9] As noted in the "Workshop 03 Action Item 21 – Reappraisal of AMS Web Site Content" report (p. 4), "Through attempting a reappraisal of the AMS Web site, it was discovered that a recent change to the Code of Procedures mandated that the minutes of the various AMS committees be posted to the Web site after approval. Subsequent correspondence with the Archivist found that: 'In addition to planning groups and commissions, all our committees and Council itself posts minutes to the website.' Although the Archivist states that such directives are internal matters and, therefore, not subject to the imposition of penalties in the event a directive is not carried out, it was thought that this may become an issue in the future" (http://www.interpares.org/ip3/display_file.cfm?doc=ip3_canada_cs09_wks03_action_21_v1-3.pdf).

[10] 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 http://www.naa.gov.au/Images/archweb_guide_tcm2-903.pdf. Last accessed April 28, 2009.

There are several advantages to the AMS in employing the Archive-It solution. Minimal input is required from the Archivist to implement the crawl process, and the Internet Archive provides hosting and storage for all archived materials. Options exist to transfer data to the institution, to enable them to store the data in addition to the storage provided by the Internet Archive.

Archive-It has been successfully implemented in many archival organizations including the University of Toronto, the Arizona State Library, Archives and Public Records, and RLG – the Research Libraries Group.[11]

The costs associated with the outsourcing option may be prohibitive in terms of financial resources for the AMS. Subscription rates range from $12,000 to $17,000 per year, however, Molly Bragg; Partner Specialist for the Internet Archive has taken a look at the AMS Web site and has offered an introductory rate of $2,000 for the first year.[12] Pricing estimates are offered for 120 days. The Internet Archive collects and stores data with minimal input from the organization, so the time saved in implementing an open source crawler and its upkeep, as well as the costs associated with data storage will be reduced.

A further issue that could become problematic for the AMS is the fact that data is stored by the Internet Archive on servers across the globe, including the USA. As the AMS is governed by the Personal Information Protection Act (PIPA)[13] it 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. Implementing a set of procedures to be followed that states what can be uploaded to the AMS Web site with strict criteria regarding personal information will ensure that PIPA is followed.

Web Site Mirroring Option

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 AMS having a Web site. We offer this solution to the AMS 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 AMS Archives wishes to preserve evidence of records appearing on the Web site.

[11] University of Toronto: http://www.utoronto.ca/; Arizona State Library, Archives and Public Records: http://www.lib.az.us/; RLG: http://www.oclc.org/ca/en/global/default.htm.

[12] E-mail from Molly Bragg to Helen Callow, April 21, 2009.

[13] Personal Information Protection Act Web site: http://www.oipc.bc.ca/legislation/FIPPA/Freedom_of_Information_and_Protection_of_Privacy_Act(May2008).htm.

Three mirroring tools were researched for the AMS. The open source crawler HTTrack and a proprietary software program "Grab-a-Site." Both have been utilized effectively in other archival institutions.[14]

HTTrack

Pros: Inexpensive; history of support; Web crawl automated; less storage requirements;

Cons: Does not address dynamic nature of the Web site; will not capture all possible user generated pages; archived site may contain broken links; minimal capture of Web site metadata will not address issues concerning presence of records on the site

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.[15] This harvester has been used successfully by archivists seeking to preserve Web content in the Microsoft / Windows environment similar to the technological environment in which the AMS operates.[16] It is thought, however, that problems seen in the remote harvesting option with possible loss of graphics and broken links could also occur if using the HTTrack Mirroring tool.

HTTrack is suggested by the author of *Archiving Websites*, Adrian Brown as a crawler option. It has been used repeatedly in the Windows environment with success. Again, the AMS must contact the Web site Content Management Provider to ensure that access is given to the spider by removing robots.txt exclusion directives and META robots tags from metadata.

Grab-a-Site

Pros: Inexpensive; history of support; Web crawl automated; less storage requirements; free trial on Web site; states ability to overcome dynamic nature of Web site (we, however, are skeptical of this)

Cons: Does not address dynamic nature of the Web site; will not capture all possible user generated pages; archived site may contain broken links; minimal capture of Web site metadata will not address issues concerning presence of records on the site; proprietary software, company unknown, uncertain if backwards compatibility offered with new software releases

Proprietary software "Grab-a-Site" from a US company called Blue Squirrel.[17] The 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

[14] 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.

[15] See the HTTrack Web site for more information: http://www.httrack.com/.

[16] For a recent discussion of implementation, see: Christopher J. Prom and Ellen D. Swain (2007), "From the College Democrats to the Falling Illini: Identifying, Appraising, and Capturing Student Organization Websites," *American Archivist* 70(2): 344-363.

[17] See the Blue Squirrel Web site for the Grab-a-Site product page: http://www.bluesquirrel.com/products/grabasite/.

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."[18] This would mitigate the presence of broken links within the copied data environment as can be the case for dynamic Web sites captured using client-side models.

This software has been implemented with satisfaction in other Archival institutions (Namely by the Electronic Records Archivist at Kentucky Department for Libraries and Archives).[19]

Adobe Web Capture Tool

Pros: Possibly inexpensive if the AMS owns a copy of Adobe Acrobat Standard (or above) (otherwise $300 (standard) or $500 for Adobe Acrobat Pro); easy to use; capture various levels of links within site; date and time stamp for captured web pages; backwards compatibility assured by Adobe; long history of support.

Cons: No metadata; reproduces flat pdf document (unable to remove portion of page to print, have to print whole page; possibly expensive ($300+); entire Web site captured each time; converts Web site to PDF rather than to PDF/A.

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.

Software converts the Web site to PDF files rather than the ISO standard PDF/A. This could be remedied by the AMS during the checks process by converting each PDF file to PDF/A to ensure long term stability. This action, however, would need to be performed manually by the Archives staff and would increase the hours spent using this product.

Acrobat 9 has limited ability to capture Flash content. It captures simple (non interactive) Flash content in a page but does not capture more complex content such as entire web pages which have been created in Flash. Media such as video on a web page is not captured.

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.

[18] Ibid.

[19] E-mail to the Management & Preservation of Electronic Records Listserv: April 3, 2009.

Comparisons

There are several technological options for the AMS to choose from in performing the desired Web crawl / snapshot; all are available at differing levels of financial and human resource inputs. Some costs are associated with all the solutions, but from differing departments of the AMS organization.

Technological Option	Used For	Price $	Human Resource Hours	AMS Office
Direct Transfer of data	Preservation of impression of content and / or Preservation of records or mandates[20]	$0[21]	5 – 10 hrs[22] per transfer; to initiate transfer; perform checks detailed below; and copy data to back-up storage media	IT Manager to perform transfer; Archives to perform checks and copying of data
Heritrix Web crawler	Preservation of evidence of records or mandates	$0	Once the Heritrix crawler has been implemented we estimate 5 hrs per transfer; to initiate transfer; perform checks detailed below; and copy data to back-up storage media	IT Manager implementation and trouble shooting; Archives to run crawl, perform checks, and backup data
HTTrack Web site Mirroring Crawler	Preservation of an impression of the Web site's content	$0	Once the HTTrack crawler has been implemented we estimate 5 hrs per transfer; to initiate transfer; perform checks detailed below; and copy data to back-up storage media	IT Manager implementation and trouble shooting; Archives to run crawl, perform checks, and backup data
Grab-a-Site Mirroring Software	Preservation of an impression of the Web site's content	$70 plus the costs associated with updates to the	Once the Grab-a-Site software is installed we estimate 5 hrs per transfer; to initiate transfer; perform checks detailed below; and	IT Manager implementation and trouble shooting; Archives to run transfer, perform

[20] As stated above, the direct transfer option could only be used to preserve an impression of the Web site content unless the procedural document that outlines the Web site update process has been implemented within the organization.

[21] Direct Transfer Pricing is made on the assumption that the AMS Information Technology Manager can implement a preservation environment for the AMS Archives at little to no cost. An additional cost may be incurred with regard to the purchase of an extra license to allow the content management system to be installed within the preservation environment. A communiqué has been sent to Whitematter to clarify any additional costs that may be associated with licensing in the preservation environment. As of this time (June 8, 2009) there has been no response from Whitematter, so this cost analysis could rise depending on information provided by the company.

[22] This is purely an estimate in terms of hours as we are unsure as to how labour intensive the direct transfer method is for a dynamic site that uses php to pull data from a MYSQL database to generate its pages.

		software	copy data to back-up storage media	checks, and backup data
Archive-It	Preservation of an impression of the Web site's content and / or Preservation of records or mandates	$2,000	If the Archive-It solution is utilized to preserve just an impression of the Web site content 2 hrs per transfer; if to preserve evidence of records a) if procedural document implemented 2 hrs per transfer; b) if procedural document not implemented much more time required	Archives to run transfer, perform checks, and backup data; if procedural document not implemented Archives to add metadata to each file (if evidence is required)
Adobe Web Capture	Preservation of an impression of the Web site's content	$0 - $300 (Standard version) - $500 (Pro version)	Once the Adobe software is installed we estimate 5 hrs[23] per transfer; to initiate transfer; perform checks detailed below; and copy data to back-up storage media	IT Manager implementation and trouble shooting; Archives to run transfer, perform checks, and backup data

Checks

Once the Web site has been captured and transferred to the AMS 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.[24]

Mandatory Requirements

Whichever solution the AMS chooses to implement within its organization, certain requirements are mandatory for all. These include file format specifications, file naming specifications, and the presence of a comprehensive retention schedule that dictates how long the data are kept.

File Format specifications

Currently, the AMS Web site contains documents that exist in a variety of formats. Budgets are usually posted in the Microsoft Excel format, minutes of meetings are found in Microsoft Word documents, and Codes and By-Laws could exist in either Microsoft Word or PDF documents. It is recommended that to ensure that the documents posted to the Web site are accessible over

[23] This time estimate would increase if the Archives wishes to convert the PDF files captured to PDF/A files.

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

time; the AMS converts all documents to a single format before posting them to the Web site. The argument for implementing a single file format is that sustainability costs are minimized when a file format of choice is built into the records creation process.

Based on our research, we would recommend that the AMS convert all files to PDF/A before posting to the Web site. PDF/A is a file format for the long-term archiving of electronic documents. It is defined by ISO as an ISO standard which was published in 2005.[25] It conforms to most of the criteria defined by Adrian Brown. The benefits of using PDF/A as the file format of choice are that it allows both PC and MAC users to access materials, although proprietary it has a long history of support, PDF allows for backwards compatibility and is the de facto standard of file formats. In addition, the Archives should retain copies of documents posted to the Web site in their original format and as a paper copy.

File Naming Specifications

The AMS should uniformly name files that are to be 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 possess 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 commas.

A suggested naming format is as follows: committee_or_group_name_name_of_document_date for example: ams_finance_commission_budget_april_2009

Retention Schedules

All data associated with the archiving of the AMS Web site should be included in retention schedules that govern the AMS'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 AMS. In addition to improved record keeping, the AMS would benefit in terms of costs associated with storage if effective disposition schedules were in place in the organization. To ensure long-tem accessibility of data it is essential that storage media is refreshed on a regular basis. If the AMS stores each iteration of the Web site indefinitely then the costs associated with refreshing media will soar over time as the data collected grows.

Procedures that Govern Web Content for Upload

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. If the procedures contain strict criteria regarding the treatment of personal information, this will ensure that PIPA legislation[26] that governs the AMS is adhered to.

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

[26] Personal Information Protection Act Web site: http://www.bclaws.ca/Recon/document/freeside/--%20P%20--/Personal%20Information%20Protection%20Act%20%20SBC%202003%20%20c.%2063/00_03063_01.xml.

Storage Options

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. Today's "state of the art" may be obsolete in 5 years time and 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[27] 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.[28] Therefore, whichever medium is chosen for storage will need to be periodically checked and/or refreshed to counteract data loss.[29]

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.[30] The Canadian Conservation Institute has put together a table that provides estimates of predicted longevity for various media storage types.

Predicted longevity of electronic media[31]

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

[27] 1995 Kodak research on their writeable CDs, reported at http://www.cd-info.com/CDIC/Technology/CDR/Media/Kodak.html, quoted a lifetime of 217 years under specified conditions.

[28] 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.

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

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

[31] Ibid.

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 AMS 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.

Throughout the AMS case study, the possibility of storing the Web site preservation data on the AMS server has been discussed with both the Archivist and with the Information Technology Manager. It has now been made clear that this is not an option,[32] so other storage possibilities have been investigated.

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 and sustainability, again documented by Adrian Brown at the National Archives of the United Kingdom.[33] 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 advises:

> 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

[32] When asked at a meeting on April 9, 2009, if the AMS server could be used as storage for the preservation copy of the AMS Web site, the IT Manager, Hong-Lok Li, replied "No, the AMS server does not have sufficient storage for this purpose. In addition, the Web server should not serve as a storage site for efficiency reason."

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

This advice will be taken into consideration.

Based on Brown's research this report will provide costs for the top three solutions for the storage of the AMS's archived Web site.

Linear Tape Open (LTO) Option

Linear Tape-Open (or LTO) is a magnetic tape data storage technology originally developed in the late 1990s as an open standards alternative to the proprietary magnetic tape formats that were available at the time. Seagate, Hewlett-Packard, and IBM initiated the LTO Consortium,[34] which directs development and manages licensing and certification of media and mechanism manufacturers. An additional benefit of the LTO technology is that it is non-proprietary, so therefore all brands of tape work in each unit. The standard form-factor of LTO technology goes by the name "Ultrium," the original version of which was released in 2000 and could hold 100 GB of data in a single cartridge. The most recent version was released in 2007 and can hold 800 GB in the same size cartridge.[35]

There are nine compliance verified licensees for LTO. These are: Fujifilm, HP, IBM, Imation, Maxwell, Quantum, Sony, Tandberg Storage, and TDK.[36] The Hewlett Packard option is described below, but the AMS could investigate the other companies for price comparison purposes.

According to the HP Web site, the HP StorageWorks RDX Removable Disk Backup System delivers an easy-to-use, affordable data protection solution for workstations and servers. Backups are simple with drag and drop file access. Long lasting removable disk cartridges and a forward and backward compatible docking station that does not require a costly upgrade for future, higher capacity cartridges, reduces costs. The system offers fast disk-based performance with the ability to store 160 GB, 320 GB or 500 GB of data on a single removable disk cartridge at speeds of up to 108 GB/hr. Portable, durable and rugged removable disk cartridges simply and securely store your backups off site for complete data protection and peace of mind.[37] The cost of such a unit ranges from $279 (US) for a 160 GB capacity machine to $729 (US) for a 500 GB capacity.[38] Cartridge prices range from $72 (US) for a single 1.6 TB cartridge to $776 (US) for a pack of 20 200 GB cartridges.[39]

[34] LTO Consortium Web site: http://www.lto-technology.com/default.php.

[35] Linear Tape-Open. Wikipedia Web site: http://en.wikipedia.org/wiki/Linear_Tape-Open.

[36] Fujifilm http://www.fujifilmusa.com/products/tape_data_storage/index.html, HP http://h18006.www1.hp.com/products/storageworks/rdx_bs/index.html, IBM http://www-03.ibm.com/systems/storage/tape/, Imation http://www.imation.com/en/Imation-Products/, Maxwell http://www.maxell-usa.com/index.aspx?id=2;14;261;574&a=info&pid=325, Quantum http://www.quantum.com/Products/TapeDrives/LTOUltrium/Index.aspx, Sony http://b2b.sony.com/Solutions/category/recordable-media, Tandberg Storage http://www.tandbergstorage.com/, and TDK http://www.tdk-media.com/professional/lto/index.html.

[37] HP Storage Works RDX Removable Disk Back-Up system: Quick Specs. Available at: http://h18000.www1.hp.com/products/quickspecs/13036_div/13036_div.pdf.

[38] HP Pricing: http://h71016.www7.hp.com/ctoBases.asp?oi=E9CED&BEID=19701&SBLID=&ProductLineId=450&FamilyId=2831&LowBaseId=21630&LowPrice=$2,499.00.

[39] Cartridge pricing single unit: http://h71016.www7.hp.com/ctoBases.asp?oi=E9CED&BEID=19701&SBLID=&ProductLineId=450&FamilyId=1455&LowBaseId=&LowPrice=&familyviewgroup=832&viewtype=Matrix; Cartridge pricing pack of 20 units: http://h71016.www7.hp.com/ctoBases.asp?oi=E9CED&BEID=19701&SBLID=&ProductLineId=450&FamilyId=1455&LowBaseId=&LowPrice=&familyviewgroup=833&viewtype=Matrix.

The system is easy to install, it is simply plugged into a USB port on the PC and the storage capacity and longevity of media is exceptional.

DVD-R Option

According to the UK's National Archives research, the DVD-R is the most effective media in terms of the AMS's needs. Gold Archival grade DVD-R has enough storage capacity to store a 4 GB Web site and is relatively affordable and easy to use. A typical DVD-R has a capacity of 4.7 GB and a cost of around $90 (CND) for a spindle of 50 units.[40] The author of the guidance note suggests using different brands or batches of the chosen media to minimize data loss due to specific manufacturers or batches having problems. The AMS should take this recommendation into consideration when purchasing media for the storage of their archived Web site, as well as the recommendation to conduct routine, periodic inspections of the files on the storage media to check for data corruption. It is also recommended that the DVD-R media be refreshed entirely every few years until testing by standards agencies has been done to discover more completely the archival capacity of the medium.

Hard Drive Option

Regarding storing the Web site on hard drives, if the AMS chooses to store their data on hard drives it is recommended that new hard drives be installed in the respective machines, or that external hard drives be purchased, so that the new hard drives can be dedicated to the archival process. As cost is an issue for the AMS, a quick breakdown of cost for various hard drives has been included in this report: 300 GB external hard drives can be purchased for as little as $70 (US) and internal hard drives range from $70 (US) for a 500 GB capacity to $95 (US) for a 750 GB capacity hard drive.[41] However, the hard drives will also need to be periodically checked and refreshed, and new hard drives purchased when the old drives reach full capacity.

Server Option

If the AMS organization has a desktop computer that is functional but not being used, it may be worth turning it into a server to store the archived Web site data. The AMS currently operates in the Windows 2003 server environment, so the Information Manager could use previously purchased software to set up the Archives' own server, or if necessary re-purchase the necessary software. Windows Server 2003 has the reliability, availability, scalability, and security that make it a highly dependable platform.[42] Pricing ranges from $199 to $999 (US) depending on the number of client access licenses—as the AMS Archives is one client, the price will be $199.

Storage Option	Benefits	Cost Financial	Cost Human Resource
Linear Tape Open	Longevity, capacity, viability, obsolescence, cost, susceptibility	$351 - $1505	Minimal: Drag and drop to copy data; perform data checks

[40] See price comparisons at the Price bot Web site: http://www.pricebat.ca/Verbatim-Archival-Grade-Gold-Ultralife-8X-DVD-R-Media-50-Disc-Spindle.p_101515/.

[41] See the New Egg Company Web site: http://www.newegg.com. Last accessed September 29, 2008.

[42] See the Microsoft Server 2003 Web site page: http://technet.microsoft.com/en-ca/windowsserver/bb429524.aspx.

DVD-R	Longevity, capacity, cost	$90	Minimal: Drag and drop to copy data; perform data checks
Hard Drive Internal	Automated, longevity, capacity, cost	$70 - $95	Minimal: automated data copying; perform data checks
Hard Drive External	Automated, longevity, capacity, cost	$70	Minimal: automated data copying; perform data checks
Server		$0 - $199	Minimal: automated data copying; perform data checks