InterPARES 3 Project

International Research on Permanent Authentic Records in Electronic Systems

TEAM Canada

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

Workshop 03 Action Item 23 – On-going Costs of Implementing Identified Technological Options

Status:	Final (public)
Version:	1.3
Date Submitted:	April 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_wks03_action_23_v1-3.pdf

Document Control

Version history			
Version	Date	By	Version notes
1.0	2009-04-30	H. Callow, E. Shaffer	Discussion draft prepared following identification of action items for CS09 at TEAM Canada Plenary Workshop 03.
1.1	2009-05-08	R. Preston	Minor content and copy edits.
1.2	2009-05-08	H. Callow	Addition of updated outsourcing pricing information.
1.3	2013-05-23	R. Preston	Minor content and copy edits for public version.

Action 23: S. Goldfarb, with assistance from the Graduate Research Assistants assigned to case study 09, to identify the on-going costs of implementing the identified technological option(s) (L. Duranti)[1]

Introduction

In Action Item 22, we researched several technological options for the Alma Mater Society (AMS) to capture and preserve its Web site data. The technological options are:

- Transfer of data from the originating source (Direct transfer)
- Remote Harvesting
    - Using Heritrix open source Web crawler
    - Using Heritrix open source Web crawler plus a log of changes to the Web site
    - Outsourcing the harvest to a third party
- Web site Mirroring
    - Using HTTrack open source mirroring tool
    - Using software designed for the purpose (Grab-a-site)

We also researched storage options for the AMS to store captured data using any one of the above capturing methods. The storage options are:

- Linear Tape Open (LTO)
- DVD-R
- Hard drive - internal
- Hard drive - external
- Server

This report will address the on-going costs of both the technological options as well as the on-going costs of storage.

As seen in Action Item 22, it is 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. We have taken this advice into consideration when providing cost estimates for the various storage media.

ON-GOING TECHNOLOGICAL COSTS

Direct Transfer

The direct transfer method involves no on-going implementation costs other than those associated with storage. Storage costs will be addressed later in this report.

[1] InterPARES 3 Project, "TEAM Canada Plenary Workshop #03: Action Items and Decisions," 4.

Remote Harvesting / Open Source Web Crawler

The open source remote harvester is free to use, so is in-line with the AMS's financial constraints.

Implementation costs associated with the open source Web crawler, Heritrix will be in terms of time spent by the Information Technology Manager in setting up the crawler and teaching the Archives how to use it. The Information Technology Manager anticipates that this could take several hours and is therefore not keen on this option as he believes open source options to be not user friendly, and due to the nature of open source, offer no support. This is not the case as the Heritrix Web page[2] offers a user manual, FAQ section and a knowledge management wiki, and is constantly updated to reflect new knowledge. It is true, however, that there is not a contact telephone number where an IT professional can be reached to troubleshoot the system.

As the Archives becomes more knowledgeable regarding the open source crawler, time spent troubleshooting by the Information Technology Manger will be reduced.

As with the direct transfer method, financial costs will be incurred by the storage options chosen. These will be addressed later in this report.

Remote Harvesting / Open Source Web Crawler / Log

The implementation costs for this option will be similar to those above. A log book can be in the form of an electronic document such as Microsoft word, or can simply be a notebook. Storage concerns are again addressed later in this report.

Remote Harvesting / Outsourcing to a third party

As seen in Action Item 22, the outsourcing option requires minimal technological expertise and can require minimal storage costs. The third party performs the Web crawl and stores the data.

The outsourcing option will require financial outlay by the AMS as it is run on a subscription basis. An introductory rate has been offered by the Internet Archive of $2,000 for the first year. For this Archive-It will perform a weekly crawl of the entire Web site and store the data on its multiple servers across the globe. Subscription costs will rise in the second year. Subscription costs range from $12,000 to $17,000; however, the Internet Archive is more than willing to negotiate these charges.

Storage will be addressed later in this report.

Web site Mirroring / Open Source Mirroring Tool

The HTTrack[3] open source mirroring tool is free to use, so is in-line with the AMS's financial constraints.

[2] Heritrix Web site: http://crawler.archive.org/.
[3] HTTrack Web site: http://www.httrack.com/.

As with the remote harvesting open source options seen above, the open source mirroring tool will initially require human resource input from the AMS's Information Technology Manager to implement the tool. Again, as the Archives become more familiar with the tool, the IT Manager's role will decrease.

Web site Mirroring / Software Purchase

The software option for Web site mirroring requires an initial financial outlay of $70. The software researched is Grab-a-Site.[4] It is a basic, straightforward program, so will require minimal effort from the Archivist and Information Technology Manager to implement. Blue Squirrel offers free trial downloads of their software, so the AMS can try out the program before making the decision to purchase.

It is impossible to predict on-going costs due to the proprietary nature of the Company, Blue Squirrel.[5] However, as the initial purchase is only $70, it is difficult to imagine further upgrades to the software being out of the AMS's price range.

Again, storage costs associated with this option will be addressed.

ON-GOING STORAGE COSTS

Linear Tape Open

Initial financial outlay for this option would be in the range of $351. This would purchase the Linear Tape Open machine required and a 1.6 Terabyte cartridge on which to store data. The entire AMS Web site currently requires 4 gigabytes of storage, so it would be many years before the AMS has to purchase additional storage space if using this option. However, it may be prudent for the AMS to purchase an additional 1.6 TB cartridge for an additional $72 to enable them to store multiple copies of the same data. This would increase the initial outlay to $524

DVD-R

Initial financial outlay for DVD-R would be in the range of $90 for a 50 unit spindle. Best practice dictates that different brands or batches be purchased to minimize data loss due to specific manufacturers or batches having problems. Therefore the costs would rise to $180.

It is recommended that the DVD-R media be refreshed entirely every two years to avoid data corruption and technological obsolescence. Therefore, the on-going costs associated with this storage method will be $180 every two years. This could be staggered to $90 every year.

Hard Drive – Internal

Initial outlay for the internal hard drive option would be $140. This would purchase two internal hard drives as it is recommended that the AMS purchase a new hard drive to dedicate to data storage as well as an additional hard drive to back-up the stored data. If a hard drive is used for storage and DVD-R is used for data back-up, the cost would rise to $160.

[4] Grab-a-Site Product Page: http://www.bluesquirrel.com/products/grabasite/.
[5] Blue Squirrel Home Page: http://www.bluesquirrel.com/.

On-going costs associated with this storage method would be $140 or $160 every two years to enable the AMS to refresh the storage media. Again, the costs could be staggered over the two year period.

Hard Drive – External

Initial outlay for the external hard drive option would be $140. This would purchase two external hard drives as it is recommended that the AMS purchase a new hard drive to dedicate to data storage as well as an additional hard drive to back-up the stored data. Again, if a hard drive is used for storage and DVD-R is used for data back-up, the cost would rise to $160.

On-going costs associated with this storage method would be $140 or $160 every two years to enable the AMS to refresh the storage media. Again, the costs could be staggered over the two year period.

Server

Depending on the licensing that the AMS already has within its organization, the server option could range from $0 to $199. The AMS already uses the Microsoft 2003 Server product, so this may be a simple and cost effective option to put in place.

An initial outlay of human resource time would be required from the Information Technology Manager to set up a server for the AMS Archives, but this time requirement would be minimal as he is already familiar with the Microsoft technology.

An additional purchase would be required of either a hard drive or DVD-R spindles to ensure multiple copies of the stored data are kept.

Server plus hard drive = $70 or $269 if the server software needs to be repurchased.

Server plus DVD-R = $90 or $289 if the server software needs to be repurchased.

On-going costs would depend on upgrades to the Microsoft server software and the purchase of additional hard drives or DVD-R for refreshing data back-up purposes.

Outsourcing

On-going financial costs associated with the outsourcing of data storage could incur depending on the AMS's choices. Data is stored by the Internet Archive and is included in the subscription cost. However, they do offer the option to ship stored data to the institution on a hard drive or electronically. This would entail a yearly additional cost of approximately $500[6] for an electronic transfer of data, plus an additional outlay of $70 or $90 for data back-up on an additional hard drive or DVD-R.

[6] Costs associated with transfer of data ranges as the fee generally "depends on how much data you archive and how we set up a transfer. We can transfer data to you either electronically or with a hard drive. Costs generally work out to $2,000 per terabyte of data transferred over disc and $500 per terabyte transferred electronically. Once you develop your collection and want to transfer the data, let's discuss costs further. I am guessing this is going to be a small collection and we can try to work with you and your budget." E-mail from Molly Bragg, Archive-It Partner Specialist to Helen Callow, May 7, 2009.