Smithsonian Institution Archives

Archiving Smithsonian Websites: An Evaluation and Recommendation for a Smithsonian Institution Archives Pilot Project

Smithsonian Institution Archives Records Management Team
May 2003

INTRODUCTION

Websites are often considered vehicles for posting ephemeral information that may have no historical value; however, the Smithsonian has since 1995 used the web to provide the public with information pertaining to Smithsonian programs, research and exhibitions. Over the past eight years, the Smithsonian has greatly expanded its presence on the web, using its website to display virtual exhibits, expeditions, and field trips; provide primary and secondary research information and educational tools; and promote involvement in Smithsonian programs and commerce through business ventures, development, and museum shop sales. The presentation of this information and much of the information itself is often unique to the Smithsonian website. Historical documentation of information found on the web was traditionally captured through paper records, but it is now apparent that historical documentation of such information will be lost if not captured electronically.

Few standards and guidelines currently exist that address the issue of long-term preservation and access to websites. Dollar Consulting was hired in 2000 to outline recommendations for capturing and preserving Smithsonian websites. Dollar Consulting has provided recommendations for preserving static HTML sites, and further research will be necessary to address the increased use of dynamic websites; however, the Smithsonian Institution Archives (SIA) can ill afford to wait for such advanced research to be completed, and will proceed in a phased approach to preserving the historical documentation of Smithsonian websites, beginning with static HTML websites. Screen prints of the original site, launched in 1995, exist, but the on-line version is lost. SIA is committed to halting the continued loss of web-based historical information.

In 2001 SIA commissioned a high-level requirements assessment for the archival preservation of Smithsonian Institution websites and HTML pages. This assessment also developed strategies, guidelines, and best practices to facilitate access to usable and trustworthy websites and HTML pages for as long into the future as necessary. One recommendation to help mitigate some of the effects of technological obsolescence was for SIA to develop a program to transfer a copy of each website and associated HTML pages to an electronic archival repository and adopt a migration strategy to repackage these pages in World Wide Web Consortium (W3C) compliant XHTML, a technology neutral format. See "Archival Preservation of Smithsonian Web Resources: Strategies, Principles, and Best Practices" (http://www.si.edu/archives/archives/dollar report.html).

Very little is known about the utility and cost-effectiveness of migration software in an on-going large-scale migration project or the resources required to implement such a program. Therefore, in 2002 SIA commissioned a follow-on study to assess the utility and cost-effectiveness of currently available software migration and validation tools and to develop a metric to estimate the resources necessary to undertake such a project. During the course of the study, TAR (Tape Archive), a technology neutral format, was explored to encapsulate migrated and validated XHTML pages. See "Archival Preservation of Web Resources: HTML to XHTML Migration Test Technical Considerations, Evaluation, and Recommendations" (http://www.si.edu/archives/archives/dollarrpt2.html).

The SIA Records Management (RM) Team reviewed the requirements outlined in Dollar Consulting's reports, and sought advice from Thomas J. Ruller (Independent Consultant for archivists and records managers working with records in electronic form), to evaluate the feasibility and requirements needed for implementing a project incorporating those recommendations.

EXECUTIVE SUMMARY

The Smithsonian Institution Archives (SIA) Records Management Team evaluated and tested Dollar Consulting's recommendations for archiving websites in late 2002. This report describes the results of that evaluation and outlines a proposal for a pilot project to archive Smithsonian websites.

Dollar Consulting provided SIA with recommendations for processing static HTML websites. However,

issues concerning the transfer, long-term preservation, and access to websites were not addressed in those recommendations. The increasing use of dynamically-generated web pages also creates a need for standards and requirements for archiving non-static websites, as well as non-textual elements such as sound, video, and digital images. Based on the evaluation and test, SIA developed recommendations for receiving, managing, and providing access to static websites; and created preliminary recommendations for archiving dynamic websites.

During the evaluation phase of the project, SIA found certain data anomalies. Specifically, the tested processes were not successful in converting poorly constructed HTML pages to XHTML. To address this issue, SIA developed draft guidelines for Smithsonian webmasters, including best practices and suggested standard metadata elements.

The overall recommendations created by SIA are included in this report, including a work flow model, and estimates for staff time required when the project is implemented.

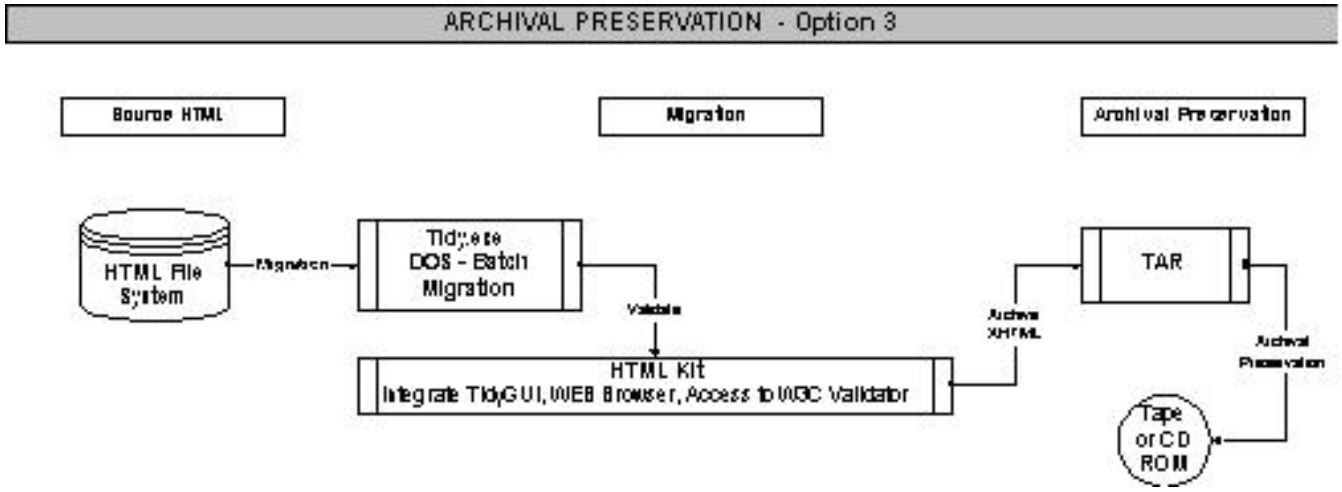
1. Evaluation of Recommendations for Project Implementation
1a. Dollar Consulting Recommendations and RM Team Evaluation

Dollar Consulting Recommendations
The RM Team tested the recommendations made by Dollar Consulting in its July 2002 paper, "HTML to XHTML Migration Test: Technical Considerations, Evaluation, and Recommendations." In this report, recommendations are made regarding migration of HTML to XHTML and encapsulation of the XHTML pages in TAR format. In its July 2001 paper, "Archival Preservation of Smithsonian Web Resource: Strategies, Principles, and Best Practices," Dollar Consulting recommended XHTML as a platform-neutral, XML compliant markup language that would be compatible with future web browsers. The TAR recommendation, first made in the July 2002 paper, would provide a means of wrapping together related web pages while maintaining their file structure and without compressing them. Encapsulated files could be saved to tape or optical media for long-term storage.

Three resources for migrating web pages were recommended. Each is based upon the same program, but provides different levels of functionality and ease of use. Tidy Utility can correct poorly-coded HTML pages and migrate them to XHTML in large, but the DOS-based program is not user-friendly. Tidy GUI is Windows-based and much more user friendly, but, although the underlying program is the same, this resource can only correct and migrate one page at a time. HTML-Kit combines Tidy GUI with a W3C Validation Service plug-in which allows direct access to the validator without needing to place the web page onto a web server. The W3C Validation Service verifies that a web page is correctly coded, following all W3C standards. Dollar recommended that SIA use Tidy Utility to save time during the migration and then use HTML-Kit to save time during the validation process.

The recommendation for TAR encapsulation was to use a DOS-based utility. The process of encapsulation and un-encapsulation using the TAR utility is non-proprietary.

The process recommended by Dollar Consulting is illustrated below.



Evaluation of Recommendations

The migration and encapsulation recommendations were evaluated using copies of the original Hirshhorn Museum and Sculpture Garden website, consisting of 31 HTML pages. At all times, a copy of the original files was kept in a separate folder to prevent irreparable changes to the web pages. The RM Team attempted to migrate a number of web pages using both Tidy Utility and Tidy GUI. The pages were then validated using HTML-Kit's direct access to the W3C validator. Finally, the RM Team attempted to encapsulate the entire website using the TAR utility and save it to a DVD+RW.

1b. RM Team Test Results and Problems and Issues with the Recommended Processes

Testbed Elements

The original testbed was 26 HTML pages and additional associated image files transferred from the Hirshhorn Museum and Sculpture Garden (HMSG) on two floppy disks in April 2002. It is unclear whether there had been a file structure at one time, but if there had been, it was now broken. Some files had also been renamed between the initial launch of the website and the time of transfer to SIA. This created many broken links between pages which had to be recreated by the RM Team. In addition, certain HTML and image files had not been included in the original transfer. The RM Team found some of the generic image files on the Smithsonian's website and the rest of the files were requested from HMSG. This resulted in a total of 31 HTML pages to be migrated to XHTML.

Migration

The RM Team first attempted migration to XHTML using Tidy Utility, a utility that "tidies" HTML pages by cleaning up coding and code nesting schemes and migrates HTML pages into XHTML. There is little documentation of the program or instructions for using it. Also, when the program was first run, files appeared to remain unchanged, and it was unclear what the error messages meant. The first test of the Tidy Utility was unsuccessful. Using Tidy GUI was successful, but slow. A second attempt at Tidy Utility with the knowledge gained from using Tidy GUI was successful and much faster than Tidy GUI.

It seems Tidy Utility is relatively simple to use, but the learning curve is large. The greatest benefit received by using the Tidy Utility is that it allows one to clean and migrate pages in batches, and it does so with great speed. After learning how to use the utility, the only major problems that occurred were due to poor coding in the testbed HTML pages, a problem further explained below.

Validation
The XHTML pages were validated using the HTML-Kit plug-in providing direct access to the W3C validator. Most of the pages were returned with multiple errors. The RM Team believes that the problems were due to sloppily-coded web pages. Many of the end tags had been missing in the original files. Earlier versions of HTML were less strict and pages could often be properly displayed by browsers without end tags. Newer versions of HTML and XHTML require end tags. Also, the coding did not follow proper nesting structures. Therefore, when Tidy attempted to logically place end tags in the document, it could not logically place the end tags in the proper sequence, occasionally violating nesting rules. Web pages with these violations could not be validated. The RM Team fixed the violations manually and re-sent the files to the validator. An added asset of the HTML-Kit software was that it could also be used as a tool to tidy and migrate individual pages that did not properly convert using the tidy utility, locating validation errors within the HTML-Kit viewer. Once a page was validated, the validator automatically added a tag to alert a browser to the variant of XHTML with which the page complied.

TAR Encapsulation
Like Tidy Utility, the DOS-based TAR utility was simple to use, but there is a large learning curve. All 31 XHTML pages with their associated image files were successfully encapsulated and saved to DVD +RW. The major problem with the TAR utility used during the test is that it follows the DOS-prescribed 8.3 file name format. File names longer than 8 characters were truncated. This truncation destroyed many links between files. Dollar Consulting had been unaware of this problem, but conducted further research and identified a Windows-based tool called PowerZip that would TAR encapsulate files without changing the file naming scheme. SIA has not fully tested this product.

2. Recommendations for Receiving Websites

Delivery methods
SIA considered several possible methods to capture and deliver websites to SIA for archival preservation. Among the possible hardware and software solutions proposed are:

1. Obtaining webcrawling software that would "crawl" the entire Smithsonian Institution (SI) website at scheduled intervals, selecting and copying only those pages/files that have been modified since the last "crawl." Although this method would capture all changes large and small, it would require substantial SIA staff time and resources.
2. Obtaining a secure read-only FTP site (file transfer protocol) from which it could issue "get" commands to access the document root of all SI web servers and thus acquire static image and page content as well as other documentation or files. The RM Team believes this option removes webmasters from the process and would undermine the SI webmaster community's trust in SIA.

Delivery Schedule
The RM Team recommends an approach whereby each SI website is captured in its entirety as a baseline. Then, all future acquisition will follow a strict, annual, staggered review in consultation with SI webmasters to determine whether an entirely new snapshot should be taken or only a partial one. This approach would ensure the capture of small changes and major redesigns, eliminate the labor-intensive operation whereby SIA staff would be forced to physically and intellectually connect individual files to earlier snapshots, and facilitate communication between SIA and SI webmasters.

The RM Team also recommends that websites and associated documentation should be transferred to SIA via CDs or DVDs. This medium was chosen because it is compact, high-capacity, widely available, and relatively inexpensive.

3. Recommendations for Archival Management of Websites and Multiple Versions of Websites

Selecting Archival Media Formats
When considering an electronic media format, issues such as cost, suitability for the project, and ease of use are considered. The RM Team chose DVD media for its large storage capacity and suitability to the project. In August 2002, the RM Team also consulted the SIA Preservation Manager to make a determination about the best DVD format for storing electronic data. DVD+RW is supported by major manufacturers including Dell and Hewlett-Packard and has been developed primarily with computing and data processing (not home video) in mind. Because of this, SIA assumes that it will continue to be successful. Of course, with formats continually changing and improving, it is incumbent on SIA to keep abreast of emergent technologies which may serve the electronic records program in the future.

File Naming Structures
When a website is accessioned into the Archives, all file names and file structures will be maintained (or repaired if necessary). An entire website will be encapsulated onto DVD and saved using a standard file naming structure created by SIA.

Tracking Accessions in CMS
Traditionally, each transfer of records into the Archives is managed through an accessioning process. Each transfer is assigned a unique accession number, and basic transfer data are entered into the Archives' Collection Management System (CMS, a collection information system (CIS)), including: who transferred the records (the office, staff person, and their title); on what date records were received and by which Archives staff; and the date the accession was acknowledged. In addition, information about restrictions on the records, the format or medium of the records, and how the records are organized is entered into CMS. CMS also records the shelf location of the records. In addition to CMS, each accession is further described and identified by the name of the record creator (office), the records title, the dates of the records, a brief accession-level description of one or two paragraphs, and a listing of accession contents (usually by box and/or folder).

For electronic records, and archived websites in particular, this information will be augmented by element-level information and metadata within and about the websites. First, information pertaining to the format and storage media is entered into an electronic records form within CMS. This includes metadata concerning the original format that was transferred to the Archives and data concerning the three archival copies of the archived website: the master copy, the preservation master, and the reference copy. Information about the original copy includes format (e.g., CD, DVD, etc.), number of media elements, the software and/or mark-up language(s) of the original website, and comments pertaining to the format or media. Information about the master, preservation master and reference copies includes format, number of media elements, the software and/or mark-up language(s) of the copies, the date the copy was made and encapsulated, the location of the media elements, and the date by which the media elements must be migrated to a new media element.

These data are used to locate the various copies of the archived website, and will also allow the electronic records program to create a tickler report showing what media elements in SIA's holdings need to be migrated.

In addition, within each archived website a folder-level document containing metadata about the transfer, conversion, and encapsulation of the website will be added to the archived website files.

Generally speaking, the master copy of each archived website will be stored off-site along with all documentation pertaining to the development, creation, and maintenance of the website. Preservation masters will be stored in a "Website Preservation Master Library," located at a second off-site facility. Reference copies will be stored in a "Website Reference Library," located on-site in SIA.

A browse list, or inventory, of the Website Reference Library will also be available at SIA for researchers to consult and browse.

4. Recommendations for Providing Access to Archived Websites

On-site Reference
Reference copies of websites will be available in one of two places for on-site researchers. Small- to medium-sized website snapshots will be placed directly onto a public access terminal. Large website snapshots will be placed on a reference CD or DVD which can be accessed on the public access terminal. Reference CDs and DVDs will be located in a "Website Reference Library" where the media will be arranged by accession number in shared boxes. The Website Reference Library will be located at SIA. Choice of location (hard drive versus optical media) will be at the discretion of the Electronic Records Program. Reference copies will be XHTML files created before the master and preservation masters are encapsulated in TAR. Additional reference copies will be created on an as needed basis by un-encapsulating the preservation master, copying the XHTML files, and re-encapsulating the preservation master.

Access to reference copies will be provided through a browse list. This list will arrange the snapshots by

museum or office and include the date taken and the scope (if only a partial snapshot). Links will be provided to snapshots that exist on the hard drive which researchers may access directly. Snapshots that exist on optical media will be followed by an accession number which the reference archivist can use to retrieve the media. Collection-level descriptions can also be accessed via Smithsonian Institution Research Information System (SIRIS) and the SIA Collection Management System. Further research must be done into the security issues surrounding researcher access to these computers, the network they are on, and the files.

Future access possibilities may involve search engines. Ideally, a search engine could be developed to search specified metadata fields, such as the creator or subject, as well as the full text of the pages. Three options for searching files exist. First, files on the public access terminal can be searched separately from each individual CD or DVD (i.e., search everything currently loaded in a specified drive). Second, all of the optical media can be placed in a juke box and everything loaded into the juke box can be searched at one time. Third, a copy of the folder-level metadata for each snapshot can be loaded onto the public access terminal, regardless of where the actual snapshot is located. The folder-level metadata can then be searched all at once and will point to the appropriate snapshot. These possibilities require further exploration.

Off-site Reference
Information about websites may be accessed remotely via the same means as other SIA collection information, including the on-line finding aids listing and search page (http://www.si.edu/archives/archives/) and collection-level descriptions in SIRIS (http://www.siris.si.edu). At this time, the RM Team prefers to retain snapshots off-line until user demands require otherwise.

Reference copies for remote researchers could be offered via one of three means. First, SIA could burn a new reference copy onto optical media and send it through traditional mail to the researcher. The second and third options are variations of each other. SIA could place the files on an FTP site or on a web server with a URL given only to the researcher. In both cases, the researcher would be given full information as to how to access and download the files and a date after which the files will no longer be available at the FTP site or URL. Reference staff should work with the Electronic Records Program staff to determine the best means of access and associated fees. This may be part of a larger reference issue concerning access to other digital files such as images.

5. Recommendations for Processing Dynamic Websites and Other Formats

Archiving Dynamic Websites (ASP sites)
Although Dollar's initial survey of SI websites indicated that 95% were static, more SI sites are becoming dynamic, integrating databases, and using content manager/delivery systems, web interfaces, and style sheets. Since content for dynamic sites is drawn from databases, Ruller recommended that SIA become familiar with these "content source" databases and gather basic information about them: their structure, function, and how the data are managed.

Although Collections Information Systems (CISs) often serve as a content source for SI's many dynamic sites, the RM Team prefers to assure preservation of CIS information in ways other than through capture as part of a website. While CISs are appraised as permanently valuable, they are also permanently active records and remain in the custody of system managers who perform archival functions on them. The only exception would be if a CIS or other permanently valuable database is in danger of destruction or will no longer be maintained, in which case it would be transferred to SIA. Instead, the RM Team recommends that SIA acquire:

- Web committee records - agendas, minutes, correspondence, design specifications, etc.,
- A snapshot of all static pages including images, audio, video and other non-textual elements, and
- Metadata, specifically
    - Site map (file structure, server locations)
    - Style sheets and ASP maps (showing from where data was pulled)
    - Creation/Update dates
    - Descriptions of the how the dynamic pages function

SIA would provide guidelines for creating metadata about pages not transferred.

Non-Textual Formats
Ruller specifically focused non-textual formats not addressed in the Dollar report. He recommended that non-textual web page elements (images, sound, video, and other multi-media) be captured and prepared for long-term preservation using standardized target formats: MPG for sound and video and uncompressed JPG for images and graphics.

Some non-textual elements can be captured and migrated to new formats easily and reliably using relatively inexpensive software tools. Image files can be captured and migrated using PaintShop Pro 6. Sound and video, however, require more complex software tools such as Adobe Premier or Sonic Foundry's Sound Forge. Of the non-textual elements, sound and video formats pose the greatest challenge for long-term preservation and access.

While image formats have tended to be stable and mostly non-proprietary (i.e., JPG), sound and video formats, especially proprietary formats such as Real Audio, will evolve. Ruller recommended moving all non-textual files to a baseline format or suite of formats to enable long-term migration and preservation.

Ruller pointed out that, over time, implementing more sophisticated software tools to manage non-textual elements will be expensive; requiring purchase of the software and training staff to use it.

The RM Team recommends additional research into these issues as an aspect of the long-term project focusing on appraisal and preservation of non-textual elements and dynamic websites. Additional study would provide the basis for establishing solid baseline formatting standards and guidelines for migration and preservation of non-textual elements and dynamic web pages.

6. Recommendations for Smithsonian Webmasters

Recommendations and Guidelines
In an effort to reduce the amount of time SIA will spend preparing each web page for migration to XHTML, the RM Team has developed draft guidelines and recommendations for webmasters. The RM Team believes that these recommendations will require minimal effort and time on the part of individual webmasters, but will provide SIA with clean, properly-coded HTML pages with basic metadata and documentation. This document, "Procedures, Recommendations, and Guidelines for SI Webmasters" (see Appendices A and B), includes best practices for linking, suggestions for validating web pages and using the resources provided by the World Wide Web Consortium (http://www.w3c.org), guidelines for developing documentation of a website, and instructions for incorporating basic metadata into each web page. These recommendations and guidelines should be used in the creation of all new web pages. This document will evolve as SIA learns more about resources available at SI and discusses best practices and willingness to cooperate with webmasters.

Future Recommendations and Requirements
SIA should seek the assistance of the appropriate staff in the Office of the Chief Information Officer for disseminating and implementing these recommendations and guidelines. The Smithsonian Webmasters Group is a potential forum for introducing this document to webmasters and receiving comments and feedback. SIA should also begin talks with Web Services staff regarding the functionalities of the Smithsonian web content management systems to determine if the systems can assist webmasters in complying with any of these recommendations and guidelines.

In the future, SIA or OCIO might establish requirements for web pages based on this document. As XHTML becomes a more widely-used standard for web page development, an additional requirement may be the migration and validation of all HTML pages to XHTML, either before they are transferred to SIA or before they are placed on the web.

7. Overall Recommendations and Estimates for Implementing the Project

Evaluation Phase Scope
This phase consisted of three tasks designed to evaluate the recommendations made by Dollar Consulting, primarily the viability of migrating HTML pages to XHTML and encapsulating these pages into TAR format, and to explore solutions to issues defined by the RM Team. The evaluation phase was to consider solutions for appraising and preserving static websites only.

The first task was to develop metadata needed to document, access, and manage web pages. The RM Team developed two documents defining metadata elements. The first document defines metadata to be placed in an individual web page (referred to as "file-level metadata"), gives instructions for entering the metadata, and identifies the party responsible for entering the metadata. The second document defines metadata to be placed in a separate document within a folder (referred to as "folder-level metadata") that would apply to all files within the folder. This folder-level metadata would be used by SIA only. Folder-

level metadata can be used as a shortcut for entering metadata for similar files within the same folder. Folder-level metadata would be used in place of file-level metadata. Within any given archived website, SIA will likely use a combination of file-level and folder-level metadata.

The second task was to archive and begin preservation of the first Hirshhorn Museum and Sculpture Garden website. This task involved accessioning the website, adding file-level and folder-level metadata, migrating HTML pages to XHTML, encapsulating in TAR the web pages and associated images, and creating a master, preservation master, and reference copy.

The third task was to enter metadata into all SIA web pages. All metadata identified by the RM Team as required from webmasters and one additional metadata tag considered optional were entered to determine the extent of effort SIA is asking webmasters to expend.

Long-term Project Scope

The long-term scope of this project encompasses all of the SI websites in their entirety. SIA will create a regular schedule for appraising all portions of the SI website and transferring full and partial snapshots of those websites. The appraisal and transfer cycle will repeat itself each year. The decision to transfer a full or partial website will be based upon appraisal criteria and discussions with the webmaster. Partial snapshots will most likely be taken when a website has not significantly changed since the last snapshot. In this case, a partial snapshot would be taken of the new or modified pages and those pages needed to put them into context.

Each transfer to SIA will be assigned a migration date on which the files will be copied to new media. A tickler system will be developed to alert staff of files that need to be migrated. The Electronic Records Program will work with the Preservation Manager to assign migration dates based upon expected media degradation and software and hardware obsolescence and to create quality control checks to be implemented between and during migrations. Part of the quality control development phase will be to review and test the recommendations made in Dollar Consulting's October 2002, report, "Archival Preservation of Web Resources: Digital Quality Assurance Tools - Technical Evaluation and Recommendations."

The long-term project will be further expanded by focusing on the appraisal and preservation of dynamic web pages and non-textual elements such as images, video, sound, and animation. This will require further research and perhaps an additional pilot program. Dynamic web pages and non-textual elements appraised as having permanent value are expected to remain physically and intellectually with static web pages obtained from the same snapshot.

An additional goal for the long-term project is to forge relations with webmasters throughout SI. This will partially be accomplished through talks with webmasters before copies of their websites are transferred. This project will only succeed with the cooperation and trust of the webmasters and two-way communications between the webmasters and SIA.

Recommendations from Dollar Consulting

The RM Team has accepted the recommendations made by Dollar Consulting to clean and migrate files using Tidy Utility and to validate files using the direct access to the W3C validator provided by HTML-Kit. The RM Team is still exploring the options for TAR encapsulation. It has rejected the initial recommendation to use the TAR utility because of the file name restrictions. The second recommendation to use PowerZip for TAR encapsulation is currently under review. The recommendation to encapsulate the original files into TAR immediately upon transfer is also likely to be rejected. The reason for doing this is to keep the original files safe in case of problems with the migration process. The RM Team feels that saving the original files in a separate folder from the copies that will be migrated achieves the same goal as the Dollar recommendation while saving time. Finally, the Dollar Consulting recommendation to save websites to tape has been rejected. DVD+RW media has been chosen instead for its stability, cost-effectiveness, and ease of use.

Recommendations from Thomas J. Ruller

The RM Team has accepted recommendations made by Thomas J. Ruller regarding obtaining a server environment on which to perform archiving work on websites. As with Dollar Consulting, the RM Team rejected the recommendation of a tape drive as a component of this server environment. A Web Library of books related to mark-up languages, operating systems and ASP technology has been created in response to Ruller's recommendations as well. The RM Team plans to further explore recommendations regarding using Teleport Pro to automatically modify internal links, integrating archival requirements into the Smithsonian's Interwoven TeamSite web content management system, and developing concrete methods for capturing and preserving multimedia components of web pages and dynamic websites. The recommendation to use a crawler to capture new and modified web pages has been rejected. The RM Team feels that the task of connecting individual pages intellectually or physically with all of the other individual pages and snapshots from a museum or office will become too unwieldy. This approach will also remove web pages from their context and will require an inordinate amount of staff time to do appraisal on a file by file basis. Ruller's alternative recommendation to obtain files through an FTP server has also been rejected because it removes individual webmasters from the process. The RM Team believes that a working relationship needs to be forged with webmasters at all levels to foster trust, cooperation, and communication. The RM Team has also rejected, at least for the near future, the recommendation to provide on-line access to archived websites via SIRIS. The RM Team believes that this could create confusion for on-line researchers and casual visitors between the current and former websites.

Division of Work

The RM Team recommends that the primary burden of appraising, transferring, preparing, migrating, encapsulating, and managing website snapshots falls to the Electronic Records Program. The RM Team will create an accession flag through which the Electronic Records Program can supply information about website transfers. The RM Team will enter the information into CMS and provide the Electronic Records Program with an accession number. The Reference Team will provide reference services on the websites with technical assistance from the Electronic Records Program. The Preservation Manager will work with the Electronic Records Program to develop adequate measures for maintaining the media and files.

The Electronic Records Program will be responsible for the following:

1. developing and managing website snapshot schedules;
2. appraising websites to determine need for full or partial snapshots;
3. negotiating transfer and means of transfer of websites;
4. communicating with and educating webmasters about guidelines, recommendations, policies, procedures, and best practices regarding issues such as HTML and XHTML code, metadata, documentation, and resources;
5. initiating, describing, and defending all new web and other electronic records projects, including bringing projects before Information Technology management;
6. researching new technologies, hardware, software, and techniques;
7. purchasing new equipment and upgrading software and hardware;
8. serving as liaison to the RM and Reference Team and the Preservation Manager;
9. and overseeing migration to new media;
10. creating accession flags in electronic form to be submitted to an RM Team liaison for data entry into the Collections Form in CMS;
11. entering all data into the CMS electronic records form after accession number is assigned by the RM Team;
12. inserting file-level and folder-level metadata into web pages and folders;
13. processing websites per the recommendations of Dollar Consulting;
14. performing migration and preservation activities;
15. creating and maintaining control over master, preservation master, and reference copies;
16. maintaining the browse list of website accessions in the Reference Library;
17. assisting reference and researchers with access to archived materials;
18. and making additional reference copies from the preservation master as necessary and documenting through reference forms what services were provided.

Time Requirements

Tests performed by the RM Team have shown that insertion of the basic metadata required of the webmasters requires approximately 1 minute per page. Metadata inserted by SIA after transfer requires approximately 1 minute per page. In consultant-timed tests, cleanup and migration using the Tidy Utility averages just under 2 seconds per page. The RM Team estimates approximately 30 seconds to validate a clean XHTML page using the HTML-Kit and about 3 hours per website for preparatory work, TAR encapsulation, documentation, and creation of the master, preservation master, and reference copy. For a small website of approximately 800 pages, completing the entire process from just after transfer to completion will require approximately 37 hours. This estimate assumes that the website was created prior to the issue of SIA recommendations and guidelines to webmasters and does not include any metadata. It is also assumed that the majority of the web pages were coded in clean HTML and do not require much manual intervention. Many websites are much larger or will not be cleanly coded.

8. Proposed Pilot

Next Steps

SIA will conduct a pilot project to test the requirements, methods and procedures for archiving static Smithsonian websites and accompanying files and images into the Smithsonian Institution Archives, ensuring the integrity, long-term preservation, and access to those sites over time. The pilot project will be built upon the initial evaluation. This pilot project will use the Smithsonian Institution Archives website, which is a larger and more complex website than the one used for the evaluation. SIA will incorporate new recommendations and processes developed as a result of the evaluation into the pilot.

The following will be used to measure the success of the pilot:

1. HTML is converted to XHTML.
2. XHTML is validated.
3. Website is encapsulated, un-encapsulated, and re-encapsulated in TAR.
4. New requirements are documented.
5. Changes to the system design and software configuration are documented.

The RM Team will oversee the pilot test, divided into three stages:

1. *Metadata Creation* - Metadata will be added to the static HTML files and additional .txt files will be created containing folder-level metadata.
2. *Conversion* - Static HTML files will be converted to XHTML files and validated against World Wide Web Consortium (W3C) standards.
3. *Encapsulation and Distribution* - The entire website will be encapsulated in TAR format. The website will be saved to DVD twice, once as a master and once as a preservation copy. An un-encapsulated copy will be saved as a reference copy. Tracking metadata about the media will be entered into SIA's Collection Management System (CMS).

Appendix A: Procedures, Guidelines, and Recommendations for SI Webmasters

As part of its mandate to preserve records of the Institution, the Smithsonian Institution Archives is responsible for maintaining copies of those portions of SI's websites that have long-term historical value. To facilitate this task, SIA requests that all SI webmasters comply with the following procedures and guidelines. Compliance on all newly-created web pages will require minimal effort and time on the part of webmasters, but will save SIA significant labor when preparing websites for long-term storage and access.

1. All links within a museum or office's website should be relative. The file path should not include folders shared with the files in which the link resides. Once a website is transferred to SIA, the folder structure will remain the same, but the root directory will change. Relative links will ensure that the link is not broken in this situation. Examples of relative links include:
   < a href="home.htm" > < a href="/John/home.htm" >

2. All links to files not transferred to SIA together, such as links to outside pages, should be absolute. The copy transferred to SIA will not be live on the internet. An absolute link will help staff and researchers find those outside pages if they still exist on the internet. An example of an absolute link would be:
   < a href=http://www.example.com/John/home.htm >
3. All web pages should contain metadata as specified by SIA. See the attached document for details (Appendix B). Most search engines do not use the metadata found on websites in their search criteria; however this metadata may be useful to staff and researchers in the future. Metadata will not affect the content or format of the web page when viewed in a browser and will only be seen by viewing the source code. Entering this metadata when the website is first created and then updating it each time the page is modified will ensure its accuracy.
4. Use the resources available from the World Wide Web Consortium (W3C) website. Useful resources include news, tutorials, and guidelines relating to HTML and XHTML; Tidy Utility download, used for cleaning sloppily-coded web pages and migrating HTML to XHTML; and the Validation Service, used for validating files as properly coded according to a specified version of HTML or XHTML.
5. All web pages encoded in HTML or XHTML should be validated before being placed on the internet. The validation process ensures that the page is coded in well-formed HTML that will be properly read by future browsers. Web pages with a URL may be validated for free at the W3C validator (http://validator.w3.org/). Some web authoring tools, as well as HTML-Kit freeware, can provide validation for pages that do not yet have a URL, often using the W3C software. Style sheets can be validated as well using the W3C's CSS validator (download at http://jigsaw.w3.org/css-validator/). Be sure to enter the validation date in a web page's metadata.
6. Webmasters should create a new site map each time the site is reorganized or on a periodic basis if pages are frequently added to the site, particularly if the folder structure does not match the site hierarchy. Previous site maps should be kept on file and transferred to SIA, along with all other documentation of the website and its development, in paper or electronic format, when the website copy is transferred. Site maps will give researchers an idea of what pages existed and how they were linked together during a particular time. This will also help SIA piece a website back together if the folder structure is ever broken.

Appendix B: Metadata Insertion

SIA has identified eight categories of metadata for webmasters to add to all newly-created websites. Some categories of metadata will not apply to all web pages, while other metadata will not change from page to page. Additional metadata will be added by SIA after a copy of the website has been transferred. This additional metadata will fulfill administrative, preservation, and access needs. The metadata scheme identified by SIA is based on the Dublin Core standard, but is more narrowly defined than the pure standard.

All metadata should be placed in < meta > tags within the < head > tag, after the < title > tag. All web pages should have a < title > tag. A metadata template will be provided. Additional metadata may be

added at the discretion of the webmaster, but should not interfere with content or format of the SIA-prescribed metadata.

Creator
The "Creator" metatag should include the full name of the museum, followed by the title of the office that created the page. The metatag can be repeated at the discretion of the webmaster to include staff members or contracting companies who contributed to creating the page.

Publisher
The "Publisher" metatag should include the full name of the museum or office responsible for making the page available to the public. Unless the page will be placed on a web server outside of the Smithsonian, the publisher will always be "Smithsonian Institution."

Date Created
The "Date.Created" metatag should include the date on which the page was coded. The date should be written in the format YYYY-MM-DD.

Date Modified
The "Date.Modified" metatag should include the date on which changes were made to the page. The date should be written in the format YYYY-MM-DD. The tag should be repeated each time the page is modified to create a running log of modification dates. Do not include additional information about the modification in this tag.

Date Metadata Modified
The "Date.MetadataModified" metatag should include the date on which changes were made to the metadata. The date should be written in the format YYYY-MM-DD. If the metadata is first created at the same time as the page, do not include this tag until changes or additions are made to the metadata. If the metadata is created on a later date than the page itself, enter the date the metadata was created in this tag. The tag should be repeated each time changes or additions are made to the metadata to create a running log of modification dates. Do not include additional information about the modification in this tag.

Date Validated
The "Date.Validated" metatag should include the date on which the page was validated as being properly coded using the W3C Validation Service, HTML-Kit, or other similar services. The date should be written in the format YYYY-MM-DD. If changes are made to the web page and it is revalidated, repeat this tag with the new date. Do not include additional information about the validation in this tag.

Format
The "Format" metatag should include the language in which the page is coded, such as XHTML, XML, or javascript. If a page is coded in HTML, this tag is optional.

Identifiers

The "Identifier" metatag should include the full URL of a page once it is live on the web. Please note that the "Scheme" for this tag in the template is "URI" and not "URL." This tag should be repeated to include the new URL should the URL change in the future. The tag should also be repeated to include a version number if the web page has been assigned one.

Language
The "Language" metatag should include the language in which the text of the page is written, preferably using the three character ISO 639-2 codes (see http://www.loc.gov/standards/iso639-2/langcodes.html for a list of codes). This tag is optional if the page is written in English.

Return to Top of Page

Contact us at osiaref@osia.si.edu

Records Management Home || Hours & Directions || Archives Division Home || SIA Home

Smithsonian Institution

Revised: May 20, 2003