Smithsonian Institution Archives

ARCHIVAL PRESERVATION OF SMITHSONIAN WEB RESOURCES: STRATEGIES, PRINCIPLES, AND BEST PRACTICES

Produced by Dollar Consulting
July 20, 2001

EXECUTIVE SUMMARY

Since 1995 when the Smithsonian Institution created its first Web site it increasingly has employed Internet technology to inform the public of various activities and programs, to offer "virtual exhibits," and to facilitate greater access to its wide ranging resources. As a result, in 2001 the Smithsonian Institution has more than seventy-five (75) Web sites and thousands of HTML pages comprising a vital component of the documentary history of the nation's leading cultural research center and museum that enriches the lives of Americans and others throughout the world. The Smithsonian Institution's use of Internet technology to carry out its mandate of the diffusion of knowledge requires a comparable commitment to the archival preservation of this "electronic corporate memory" that these and future Web sites and HTML pages represent. Without such a commitment this "electronic corporate memory" will disappear and future generations of Americans will be deprived of the opportunity to view, understand, or appreciate the vital role of the Smithsonian Institution in the diffusion of knowledge in the late 20th and early 21st centuries. The need for such a commitment to the archival preservation of Smithsonian Institution Web sites and HTML pages is heightened by the fact that an exhaustive search failed to locate an extant, usable copy of the first Smithsonian Web site that was created in 1995.

The Smithsonian Institution Archives commissioned this study to conduct a high-level assessment of requirements for the archival preservation of Smithsonian Institution Web sites and HTML pages and to develop a strategy, guidelines, and best practices that would facilitate access to usable and trustworthy Web sites and HTML pages for as long into the future as may be necessary. The study consists of five chapters and two appendices that examine the following five topics:

- Archives and technology framework that affects the archival preservation of Web sites and HTML pages,
- Overview of the Smithsonian Public Web server and the Web sites and HTML pages of three museums (NASM, NMAH, and ASIA),
- Appraisal of Web sites and HTML pages for long-term retention,
- Recommendations and guidelines,
- Documentation of Smithsonian Institution Web sites and HTML pages, and
- Preservation metadata model.

The archives and technology framework that affects the archival preservation of Web sites and HTML pages involves three different components. The first is the records life cycle management of Web sites and HTML pages, which addresses long-term access considerations that must be taken into account during their design, use, and maintenance. For example, the use of non-standard HTML markup or reliance upon proprietary software can create major impediments to long-term access. Another impediment is the failure to recognize at the outset that Web sites and HTML pages must be preserved to ensure the protection of the corporate memory of the Smithsonian Institution. The second component of this framework is a discussion of sixteen Web technology tools that include the various versions of Hypertext Markup Language (HTML), static versus dynamic Web sites and HTML pages, third-party Web sites, and Web search engines and indexes and how they can affect the archival preservation of Web sites and HTML pages. The third component is an examination of three electronic archiving strategies: data archaeology, emulation, and migration.

A viable electronic archiving strategy for preserving Smithsonian Institution Web sites and HTML pages must take into account the public Web technology infrastructure and various practices that exist in more than seventy-five Web sites. A detailed examination of each of these sites was beyond the scope of this study so the Web sites and HTML pages of three museums - the National Air and Space Museum, the National Museum of American History, and the Freer Galley of Art and the Arthur M. Sackler Gallery of Art - were examined in some detail. Special attention was given to mapping the National Air and Space Museum Web site and HTML pages to archival appraisal criteria. Based upon the findings of these three museums it is likely that most Smithsonian Web sites and HTML pages have been upgraded to version 4.0 of HTML and that few dynamic HTML pages now exist. The latter is significant because it suggests that in the short run, say, the next five years or so, the Smithsonian Institution Archives can concentrate on the preservation of static Web sites and HTML pages where the preservation problems are well understood and manageable.

Archival appraisal is the process of selecting Smithsonian Institution Web pages and HTML pages for long-term retention. Prior to actually conducting this archival appraisal two questions must be answered: (1) are Web sites and HTML pages records? and (2) what is the unit of appraisal? This study argues strongly that Smithsonian Institution Web sites and HTML pages are records because they serve as evidence of how the diffusion of knowledge was carried out. The unit of appraisal, that is a body of information that is dealt with as a collective entity, is called a Web records series, which consists of Web

source material under the control and management of a Smithsonian entity and is "owned" by the entity that is primarily responsible for its creation and maintenance. Consequently, Web sites and HTML pages hosted on third-party sites that are not under the control of the Smithsonian Institution in fact are not Smithsonian records.

The study concludes with six areas of recommendations for follow-on activity that reflect the records life cycle management approach.

1. The Smithsonian Institution Should Adopt A Web Policy That Mandates:
    ❍ The life cycle management of Web-based records in which a Web site is defined as a record series,
    ❍ Archival capture of Web sites and HTML pages on an archives Web server,
    ❍ Minimization of software dependence through the use of vendor technology neutral formats,
    ❍ Minimization of hardware dependence through periodic transfer of SI Web sites and HTML pages to new storage media, and
    ❍ Capture of audit trail metadata about preservation activities taken to extend the usability of Web sites and HTML pages.

2. The Smithsonian Institution Should Incorporate Long-Term Access Considerations in the Design and Authoring of Smithsonian Web Sites and HTML Pages That Require:
    ❍ Use of Dublin Core Metadata Elements (or their functional equivalents) for search and retrieval,
    ❍ HTML markup to be XML compliant,
    ❍ Avoiding the use of a third-party proprietary search engine that is not under the control of a SI Webmaster,
    ❍ All links with the same server to be relative (not absolute),
    ❍ Periodically copy and maintain Web sites and HTML pages, and
    ❍ Major revisions to a Web site to be fully documented.

3. The Smithsonian Institution Archives Should:
    ❍ Use the definition of a Smithsonian Institution Web records series that has a unique URL whose content is created and maintained (i.e., "owned") by an identifiable Smithsonian organizational entity,
    ❍ Generate a Smithsonian public Web site map and identify each records series,
    ❍ Generate a Web site map for each Smithsonian Institution Web records series and identify the HTML pages that belong to each records series, and
    ❍ Schedule each Smithsonian Institution Web records series for permanent retention and transfer to the Smithsonian Institution Archives electronic archives repository.

4. The Smithsonian Institution Should Capture Web Sites that:
    ❍ Consist of a snapshot of each Smithsonian Institution Web site and associated HTML pages as a baseline,

○ Establish a history change log of each Smithsonian Institution Web site and document all revisions,

○ Capture a new snapshot of any new HTML pages added to a Web site, and

○ Capture a new snapshot of a Web site when it is redesigned, there are major changes, or a new technology platform is utilized.

5. The Smithsonian Institution Archival Preservation Strategy Should Entail:

○ Adoption of the Digital Archaeology Strategy as a means of acquiring immediate custody of Smithsonian Institution Web sites, HTML pages, and associated metadata (and documentation),

○ Transfer of Smithsonian Institution Web records series that is not overly complex to the electronic archives repository so that archives staff may obtain hands-on experience working with a Web records series, and

○ Implementing the Migration Strategy over the next two to four years as resources are made available.

6. The Smithsonian Institution Should Adopt Preservation Principles and Best Practices that Involve:

○ Creation of a Smithsonian Institution Electronic Archives Repository to serve as a trusted third party,

○ Transfer of Web sites and HTML pages to the Smithsonian Institution Electronic Archives Repository,

○ Ensuring the continuing processibility of Smithsonian Institution Web sites and HTML pages,

○ Preserving the content and contextual integrity of Smithsonian Institution Web sites and HTML pages,

○ Employing XHTML as a vendor technology neutral format,

○ Selection of a durable and cost effective off-line storage medium,

○ Maintenance of a secure electronic archives repository,

○ Implementation of a quality assurance media monitoring program,

○ Migration of Web sites, HTML pages, and associated metadata as often as necessary to avoid technology obsolescence,

○ Creation and maintenance of an historical preservation log for Smithsonian Institution Web sites and HTML pages.

---

1. INTRODUCTION

1.1 Purpose

This report provides the Smithsonian Institution Archives (hereafter SIA) with a set of recommended guidelines for the archival preservation of Web sites and HTML pages. The purpose of the guidelines is to help the SIA implement appropriate strategies and procedures that ensure the capture, management,

and preservation of Smithsonian Institution (hereafter SI) Web sites and HTML pages for as long into the future as may be required.

The framework for this study incorporates an integrated records life cycle process model that links the design and maintenance of SI Web sites with their archival appraisal and long-term preservation and usability. Another key component of this framework is an understanding of potential technical issues associated with SI Web sites and HTML pages that can create difficult, and in some instances intractable, barriers to their long-term preservation as complete and accurate historical resources.

1.2 Challenge

Senior SI officials are encouraging various offices, museums, and research programs to expand their use of the Internet to inform the public of various activities and programs, to offer more "virtual exhibits," and to facilitate greater access to their wide ranging resources. As a result, since 1995 when the SI created its first Web site, there are now at least seventy-five and thousands of HTML pages. There are no overall policies, procedures, and systems in place to ensure that trustworthy (1) and usable documentation of Web-based activity is created and preserved. If this situation is not corrected then the SI will lose a significant part of its "electronic corporate memory" over time and fail to meet legal and institution obligations and community expectations. From a broader perspective, SI Web-based activities constitute a vital component of the documentary record of the nation's leading cultural research center and museum that enriches the lives of Americans and others throughout the world. The failure to ensure the preservation of this "electronic corporate memory" will deprive future generations of Americans of the opportunity to view, understand, and appreciate the vital role of the Smithsonian Institution in the diffusion of knowledge in the late 20th and early 21st centuries.

The SIA mandate includes the task of working with all components of the SI to help ensure that these legal, operational and historical imperatives are met. This report is intended to assist the SIA in meeting this responsibility.

1.3 Scope

A study of the archival preservation of SI Web sites would involve ideally detailed interviews with SI webmasters and the compilation of an inventory that lists key information about each Web site and HTML page. However, the public SI Web server hosts or has links to more than seventy-five Web sites that aggregately contain thousands of HTML pages. (2) The resources required to compile such a comprehensive inventory were not available. Instead the Project Director decided to conduct a general review of the most prominent SI Web sites and to focus in some detail on a smaller number of SI Web sites. Accordingly, the National Air and Space Museum (NASM), the National Museum of American History (NMAH), and the Freer Gallery of Art and Sackler Gallery were selected. These three museums are likely to contain most of the types of HTML pages and associated technology issues found in Web sites across the SI. (3) The National Air and Space Museum Web site will be reviewed in greater depth and mapped against SIA appraisal criteria.

1.4 Report Organization

This report consists of five chapters and two appendices. Chapter 1 comprises an introduction to the study and to the concept of life cycle management of Web sites and pages, which is the underlying premise of the study. Chapter 2 examines the archives and technology framework that affects the archival preservation of SI Web sites and HTML pages. An overview of the public SI Web server and the Web sites and pages of three museums is the focus of Chapter 3. Chapter 4 reviews issues associated with the archival appraisal and retention of SI Web sites and pages. Chapter 5 offers a number of recommendations regarding a SI Web preservation policy, the design and authoring of SI Web sites and HTML pages, archival appraisal of Web sites and pages, archival capture of SI Web and HTML pages, a Web archival preservation strategy, and SIA Web site preservation best practices. Appendix 1 contains recommendations on Web site documentation (metadata) that should be captured at the time of transfer to an electronic archives repository. Appendix 2 consists of a recommended Metadata Preservation Model for the SIA to follow in ensuring the archival preservation of usable and trustworthy SI Web sites and pages.

2. ARCHIVES AND TECHNOLOGY FRAMEWORK

2.1 Records Life Cycle Management Model

The life cycle of records is generally understood to mean the life span of records from their creation or receipt until their disposition. This life span can be characterized as creation (receipt), use and maintenance, and disposition through destruction or transfer to an archives repository where access and preservation activities are conducted. The life cycle of paper records was sequentially structured along a time continuum that consisted of three stages: active, semi-active, and inactive. The active and semi-active stages tended to be associated with creation and use and maintenance activities under the purview of records management while the disposition stage was associated with storage in an archives repository or destruction under the purview of archives.

The advent of electronic records quickly demonstrated to archivists and records managers the ineffectiveness of this model. Pre-custodial intervention to rescue electronic records that were being inadequately cared for became the major focus of the electronic records program of the National Archives of the United States in the 1970s and 1980s. (4) Rescue gave way to a more systematic intervention to ensure that electronic records were being properly created, managed, and documented and to identify those that merit long-term retention. It became increasingly evident that not only must these electronic records of long-term value be identified, but that provisions for ensuring their persistence until they are transferred to an electronic archives repository must be in place. (5) The design, creation, and maintenance of usable and trustworthy electronic records could not be left to chance or the good intentions of records creators and computer information specialists. The lack of

understanding and/or the absence of professional recordkeeping experience during the system design and creation phases of records can lead to "record disasters" that are not understood or recognized until it is too late to do anything. As Adrian Cunningham of the National Archives of Australia puts it:

> Failure to pursue a more active agenda will leave us patiently waiting at the railway station for the goods train of life to deliver the unreliable electronic leavings . . . . Years of passive and patient pacing of the platform will come to an end, I fear, when the whistle blows and the train pulls into the station and we finally come to the full realization that it is full of ghosts and that ghosts do not satisfy our researcher's need for solid, reliable, and authentic evidence of the past. (6)

One example of this is the point in time when the retention of electronic records is scheduled. It is universally acknowledged that the assignment of records disposition instructions (i.e., how long to be retained before destruction or transfer to an archives for long-term retention) should occur at or shortly after the time of creation and certainly no later than transfer to a recordkeeping system. The life cycle management of electronic records known to be of long-term value (a retention period of ten years or longer) requires that additional steps be taken early on to facilitate long-term access. Among the steps that merit consideration are:

- Use the Dublin Core Metadata Element Set (7) (or its functional equivalent) to assign descriptive elements to Web sites and HTML pages to promote uniform and consistent searching and retrieval,
- Minimize software dependence through the use of vendor technology neutral file formats,
- Minimize hardware dependence through the transfer of electronic records to new storage media for which there is a strong market presence, and
- Capture of audit trail metadata about preservation activities taken to extend the usability of electronic records.

The failure to take these steps early in the life cycle management of electronic records can, at best, give rise to intractable problems in the future or to problems that can only be rectified at a great expense or, at worst, render the electronic records unusable.

2.2 Web Technology Issues That Affect Archival Preservation

Since its inception in the early 1990s, the World Wide Web (W3) has undergone dramatic change as new authoring tools, techniques and browsers were introduced that supported the creation and use of Web sites and HTML pages. Some of these tools, techniques, and browsers were proprietary and were short-lived while others have persisted in one form or another. In addition, some authors of Web publications began to combine HTML tags in non-standard ways that, while very effective in the short-run, created long-term access issues. Over time some of these tools, techniques, browsers, and non-standard use of HTML tags have become impediments to archival preservation. This section briefly reviews Web technology issues that include the various versions of HTML, dynamic versus static

HTML pages, and third party Web sites and HTML pages to promote a better understanding of how they may affect the archival preservation of Web sites and HTML pages.

2.2.1 Hypertext Markup Language (HTML)

HTML is considered the "lingua franca" for publishing hypertext on the Web. It is a non-proprietary, vendor technology neutral standard based on Standard Generalized Markup Language (SGML) that was developed to support the sharing of textual and graphical information on the Internet without regard for the hardware or software used by authors and users. HTML uses tags (e.g., < h1 > and < /h1 >) to structure text into headings, paragraphs, lists, hypertext links and the like. The first version of HTML, 1.0, was released in March of 1993 and supported a limited number of "tags" and functions. Since its release there have been four versions: 2.0 in 1994, 3.2 in 1996, 4.0 in 1998, and 4.01 in 1999. (8) HTML 2.0 established tags and functions that have become common to all HTML documents and browsers. HTML 3.2, 4.0, and 4.01 support the tags and functions in HTML 1.0 and 2.0. This backward compatibility feature of HTML through version 4.01 ensures that any browser compliant with HTML 4.01 can render HTML documents.

2.2.2 Extensible Hypertext Markup Language 1.0 (XHTML)

The future of HTML seems assured because the W3C recently announced a new version of HTML that reformulates HTML 4.01 as an XML application. Called XHTML, it combines HTML markup tags for vector graphics, math, and E Commerce, which means that as an XML application it will run on different platforms and help ensure a consistent rendering of HTML content across different computers. XHTML will be available in three "flavors": Transitional, Strict, and Frameset. The most important of these is XHTML Transitional because it incorporates a style sheet feature that helps ensure consistent rendering regardless of the browser. Old HTML (2.0 - 4.01) can be reformatted as XHTML using a non-proprietary software tool developed by W3C called "Tidy utility." It will correct markup errors and non-standard use of the Font tag. (9)

2.2.3 Browser dependent HTML tag extensions

This situation occurs when a tag used in one browser is not supported in other browsers. In its efforts to provide greater flexibility to users, Netscape provided a number of HTML tag extensions, especially to HTML 2.0, (10) that work with Netscape browsers but are ignored by other browsers such as Internet Explorer. For example, earlier versions of Netscape supported a < Center > tag to define a section of text that would be centered in windows rendered by Netscape. Browsers that do not support the < Center > tag will ignore it and align the text from the left margin. (11) The presence of HTML tag extensions in an HTML document means that it will have the same look only on a browser that is compatible with the one used to create it. When rendered on a browser that does not support specific HTML tag extensions, a document can look substantially different even though the informational content is identical.

2.2.4 Browser idiosyncrasies

There are substantial differences in browsers that may affect the rendering of Web material. One reason why two different browsers may render or display Web material differently is due to type fonts. In some instances, a type font that was available to the author of Web material may not be available to the end user or viewer. Therefore, the end user's browser will employ its default type font. HTML filters from word processing applications and some HTML authoring tools may generate code that attempts to identify the physical layout of documents when they are rendered by combining the tags Font, BR and (not-breaking spaces). Some authors also used $<p></p>$, which represents a blank paragraph, to add vertical white space. A related issue is the use of color to call attention to certain material. Differences in type font and color may seem insignificant, but they do affect the look and feel of Web based records. Such differences in Web browsers, it should be noted, do not affect the content of Web sites and HTML pages. More importantly, as XML and XHTML become firmly established organizations will upgrade their sites and HTML pages to these standards so for all practical purposes these differences will only exist in legacy Web sites and HTML pages.

2.2.5 Style sheet

HTML is based upon the Standard Generalized Markup Language (SGML) that makes electronic documents software and hardware independent so they can be processed on any modern computer platform. This is accomplished by separating the logical structure of documents from their physical structure. SGML tags ($<>$) are embedded in documents to denote such things as headings, paragraphs, lists, tables, and the like. The physical layout of a document, which includes type font, margins, spacing, etc, is determined by layout instructions called a style sheet that is applied to the logical structure when a document is rendered on a monitor or printer. In the early years of HTML, some users chose to embed layout instructions for the physical structure along with tags for the logical structure. Subsequently, W3C developed an HTML style sheet called Cascading Style Sheets (CSS) that are parallel to style sheets for SGML. The HTML Tidy utility, mentioned earlier, can clean up some of these layout problems by converting them to CSS.

2.2.6 Relative and Absolute addressing

These two terms refer to two different ways to establish the path of a URL that leads to a specific folder or HTML page. An absolute URL address means that the location of a file is given in terms of its location vis-à-vis the root directory and could be displayed as //Dir1/Dir5/Dir6/file.ext. A relative address means that the location of a file is relative to the location of another file and typically is displayed in a URL as ../Dir5/Dir6/file.ext. A relative address is preferred because if the directories containing HTML folders and pages are moved intact to another computer or server the links are unlikely to be broken. In contrast, when absolute addresses are used the links to files in a directory (or directories) are likely to break and reestablishing these links can be a time consuming task if many files are involved.

2.2.7 Static HTML Pages and Web Sites

An HTML page, which is sometimes called a Web page, is information marked up in HTML that contains text and specifications about where images and other multimedia material are to be placed when the page is displayed. An HTML page may contain a single image or multiple images. An HTML page also may contain text that fills only one page or extends over multiple pages. Therefore, the concept of pagination in printed pages is absent in HTML pages. Static HTML pages, which may contain only text, images, or a combination of multimedia, display the same content and layout ("look and feel") to all viewers. By clicking on hypertext links, viewers can follow pre-established navigation routes from part of one HTML page to another part in the same page, from one HTML page to a different HTML page, from one HTML folder to another folder, or from one part of the site to another site.

A Web site is a collection of HTML pages residing in one or more folders on a server and tied together with hyperlinks in a hierarchical structure. At the top of this structure (or root) is a folder called a "home page," which is the first thing that viewers see. A "home page" displays hyperlinks to all of the HTML folders that comprise the Web site and to any third-party Web site or folders that share a common address. For example, the Uniform Resource Location (URL) for the Smithsonian Institution home page is **"http://www.si.edu"** while the URL for a third-party Web site in the NMNH Web site is **"http://seawifs.gsfc.nasa.gov/ocean_planet.html."** A static Web site displays the same information to all users and, as noted above, its only interactive feature is "clickable navigation." Static Web sites can be captured as "snapshots" at a given point in time and links to third-party sites will remain active until they are closed down. However, over time the view users have of third-party Web sites that remain active or open will reflect its current status in which HTML folders and pages may have been updated, replaced, or removed.

2.2.8 Active Server Page (ASP)

An Active Server Page is an HTML page with embedded executable code written in what is called a scripting language and saved with the .asp extension. When a user requests a page that is an active server page, the Web server navigates through the page looking for executable ASP code. After the code is run, the ASP portion of the page is stripped out and a complete HTML page is sent to the browser. The actual content of the HTML page would be determined by a set of rules or procedures embedded in the ASP that are invoked based upon responses users make to a series of questions. Like a database as discussed below, HTML pages may appear to be static because a display template is used for viewing but in fact the content of these page exists only at the time the query response is presented to a viewer. In this regard, Active Server Pages are similar to software applications. (12)

There are two ways to ensure the preservation of the output from Active Server Pages, assuming of course that the output merits long-term retention. One way is to capture a copy of each HTML page that is generated from an Active Server Page and set it aside as evidence of the transaction. The other way is to capture the rules or procedures embedded in the ASP code and the responses or queries from users so that they could be executed against the system in the future, which must be fully operational.

2.2.9 Data Streaming

The delivery of multimedia data (audio and video) over the Internet is a rapidly growing area of Web-based communications. The large amounts of binary representation that multimedia data such as audio and video require increase the amount of time for data transmission. Data streaming allows Internet users to immediately display the information rather than waiting for an entire file to be downloaded. In effect, data streaming allows for the flow of information to occur in a continuous fashion much like a stream of water from a water faucet. Without data streaming users would have to wait until an entire file is downloaded to a receiving computer before the information can be displayed.

Data streaming, typically audio and video streaming, in an Internet context is particularly relevant when real-time audio and video are involved. An example of data streaming is the "broadcast" of audio and video signals over the Internet as in the case, for example, of NBCI.com. The archival preservation of "data streaming" Web-based sources is especially challenging for two reasons. First, it requires the capture of the streamed data as it occurs in a real time environment. Second, streamed Web-based sources are audio and video whose binary representations typically are voluminous and even when compressed require high capacity storage media. (13)

2.2.10 Third-Party Web sites and HTML pages

Frequently a Web site or an HTML page will have one or more hyperlinks to other Web resources on servers that are external to the one hosting the Web site in use. The fact that these Web resources are under the control of a third-party means that their owners can undertake any number of activities, ranging from changing technology platforms to closing down a site or page, that result in a Web site and/or HTML page where the hyperlink cannot be established. Many Internet users are familiar with the message declaring that a site or page cannot be opened. The vulnerability of third-party Web sites and HTML pages is likely to increase greatly with the passage of time as they are closed or are modified. This vulnerability can be mitigated by downloading a copy of these Web sites HTML pages to the archives Web server at the time of transfer. But this solution has significant consequences. First, it will increase the volume of data that must be preserved, which increases long-term storage costs. Second, if the third-party Web site uses proprietary software tools such as a search engine, a database, or file formats then it will be necessary to migrate them to new versions as they become available. Absent such migrations, at some point in the future the tools will become absolutely obsolete and no longer can be executed.

2.2.11 Version control

Charles McClure and Tim Sprehe call for the creation of an historical log or audit trail for each Web site (and subsite as appropriate) that captures a time-line of events and activities associated with management of the site. (14) This time-line can provide an historical context that answers who, what, and when questions about changes in a Web site. Clearly, from an archival perspective this version control or audit trail of information is part of the metadata that is essential to a full understanding of the

Web site and its content. If such a version control log or audit trail is part of a Web site it should be transferred at the same time as the Web site is transferred and incorporated into the HTML directory supporting the site.

2.2.12 Plug-ins

A Web plug-in is a separate software application or module that is automatically invoked by an HTML tag (Embed) without users having to open the same application from a pick list or menu. Typically, plug-ins offer features and flexibility not provided by a specific browser. Widely used Web plug-ins include Real Player, Live3D, Macromedia Shockwave for movies and animation, RealAudio for audio, and Adobe Acrobat (Amber) for rendering PDF documents inside a browser. Plug-ins can create impediments to archival preservation because they usually are native to the platform on which the browser runs, which means it might not be possible to replicate the functionality on another technology platform. In addition, plug-ins are native software applications that require vendor support and updates as technology changes. Without on-going vendor support over many years, it will be difficult, if not impossible, to fully reconstruct all of the multi-media functionality of a Web site or an HTML page. Currently, there is no archival preservation solution for Web plug-ins although it is possible that vendor technology neutral Web plug-ins will be developed. The only viable archival preservation alternative is to document that a Web plug-in, which was part of the original Web site or HTML page, no longer is available. This documentation should include a brief description of the function the plug-in performed and where it was used.

2.2.13 Databases

A growing Web technology trend is to use a Web site as a front-end to an organization's database. Viewers may search prepared topics or menus that link to the content of a database or they may frame their own queries that are run against the content of a database. Typically, the search results are retrieved and displayed as an HTML page. In some instances, the displayed document may exist as an entity instance in a database so it will have a unique identifier (URL). In other instances, the actual HTML pages are generated on the fly based upon user preferences, business rules, or the capabilities of a Web browser. The content, structure, and presentation of an HTML page generated on the fly may appear to be static because a display template is used for viewing but in fact the HTML page exists only at the time the query response is presented to a viewer. (15) It is possible to reconstruct such HTML pages if the display template, user preferences, and business rules are retained and run against the database. However, this becomes problematic when the database is updated frequently. In the long-run the most critical aspect of the use of a database in Web sites is that search engines are proprietary products and unless the data is stored in a vendor technology neutral format it will be difficult to transfer the data from the old database to the new one.

2.2.14 File Transfer Formats

The function of a file transfer format is to facilitate the transfer of electronic data/records from one

technology platform to another. In terms of Web-based records, there are two file formats that may be used. One is the File Transfer Protocol (known generally as FTP), an Internet "de facto" standard that is used in an "on-line" environment to download or upload electronic data/records from one application platform to another one that either is similar or different. When using FTP, the Target platform, that is, the new application, is known.

The second file transfer format is designed largely for use with removable storage media, such as magnetic tape or CD-ROM. It is called Tar (Tape Archive) and initially was designed to function with UNIX to aggregate a number of specific files into a single file and write the file onto magnetic tape for storage or input to another application platform. Tar utilities are now available to support Unix, Windows NT, and Novell. (16) There are two distinct advantages to using Tar. First, it can capture the directory structure that supports a Web site. Second, unlike FTP it is not necessary to know the Target platform at the time Tar files are created, which is particularly useful for secondary storage and long-term preservation.

2.2.15 Web Snapshots

A Web snapshot should be a complete, accurate, and executable copy of a Web site (or HTML pages as appropriate) at a particular point in time that replicates all of the relevant functionality of the site so that it can be "reconstructed" in a different computer environment. (17) In effect, a Web snapshot is an historical documentary source that has been frozen in time. A Web snapshot, it should be noted, retains links to Web sites, including "plug-ins" that must be downloaded from a third-party source, such as a software vendor. Unlike Web snapshots, some third-party Web sites are likely to be modified over time so viewers will be able to access only the most recent version of the sites. Furthermore, over time these links may be closed down or substantially upgraded. Any link or functionality in a snapshot that cannot be reconstructed should be fully described in metadata that comprises part of the preservation history of the Web site. As noted earlier, one potential solution would be to download all third-party Web sites and incorporate them into the Web site and HTML pages that are transferred to an electronic archives repository.

A Web snapshot may be "mirrored" or replicated on an electronic archives repository through the use of FTP or written in Tar to magnetic tape or CD-ROM for off-line archival storage. Both the National Archives of the United States (18) and the National Archives of Australia (19) have advocated the use of snapshots to capture Web sites for transfer to an archival environment.

2.2.16 Web Search Engines and Indexes.

A Web search engine is a database system designed to index Web sites and HTML pages and to link user entered search terms to this index to retrieve documents that contain these search terms. A typical Web search engine has a special program, usually called a spider or crawler, that goes to a Web site, and following the various links (like a spider web), navigates the entire site and then retrieves all of the HTML pages and images that are found. Later the search engine will analyze the data and extract words

and HTML page links that are used to create an index to the Web site. Some Web search engines are included as part of a Web authoring tool such as ColdFusion (Verity search engine) so the index creation is confined to a collection of documents within ColdFusion. Web search engines are proprietary so a given index is usable only with the search engine used to create it. An added layer of complexity may occur when the index is created by a proprietary search engine at no cost to the owner of the Website in exchange for allowing the index to be hosted on the vendor's platform. The Website owner has little or no control over the index, including the frequency of updating.

A Web index to HTML pages is not static because as new pages are added or as viewer preferences are identified new index terms are added. For example, each week a Web crawler navigates the NMAH Website and adds new index information. This means that a snapshot of a Web site will reflect the index as of the date of the most recent crawl.

The proprietary nature of Web search engines means that a snapshot of a fully functioning Web site must include a copy of the search engine because an index without the search engine (or vice versa) is useless. This may require a user license for the search engine to ensure access to the search engine and index. In the long run, maintaining a searchable index of a Web site captured in a snapshot becomes problematic as search engines become obsolete. The problematic nature of continuing access to a searchable Web index increases if the index is maintained on a third-party server that also hosts the Web search engine.

2.3 Archival Preservation Strategy Options

Ensuring the accessibility and usability of trustworthy SI Web sites and pages as evidence over time faces three fundamental impediments: the limited useful life of storage media, a constantly changing technology that inevitably creates technology obsolescence, and software dependency. The combined consequence of these three impediments is the inevitable loss of long-term access to useable and trustworthy Web based records. Currently, there are three different strategies about how to deal with these consequences:

- Data Archaeology
- Emulation
- Migration

2.3.1 Data Archaeology

Seamus Ross of the University of Glasgow advocates one approach, which is called Digital Archaeology. (20) Essentially, Digital Archaeology is a minimalist approach whose primary focus is to move electronic records in their current format and software dependency into a recordkeeping environment and to convert them to a new technology only when future access to them is required, however long that may be. During this period the electronic records will remain executable for as long as the original application software and operating system are operational. Of course the electronic

records can be viewed for as long as a viewer is available that can interpret the records but at some point in time a viewer with this capability will not be available. The only way now known to convert non-executable electronic records to executable electronic records is to write computer code or programs, which is very labor intensive and costly.

The Digital Archaeology Strategy transfers the cost of converting non-executable electronic records to executable electronic records to an undefined point in the future. The Digital Archaeology Strategy is not without cost because it requires maintaining the continuing readability of electronic records by periodically transferring them to new storage media. In addition, this strategy requires collecting and preserving documentation about operating systems and application software or having access to such documentation that would be used in the future to replicate the technology platform on which the electronic records were originally created, used, and maintained. This facet of the Digital Archaeology Strategy can be characterized as future reverse engineering. (21)

2.3.2 Emulation

The second strategy is known as Emulation and is closely associated with Jeff Rothenberg of the Rand Corporation. Rothenberg defines "Emulation as a process in which one computer is used to reproduce the behaviour of another computer with such fidelity that the emulation can be used in place of the original computer." (22) Emulation supports what Rothenberg calls "executable digital originals" that are retained in their original native processing format (e.g., WordStar) that can be processed only with specific hardware, operating system, and application software. Emulation requires the development of a computer program that replicates all the functionality of the hardware and operating system. This Emulator, which would require some updating over time as technology changes, would be used to replace the original software application and operating system used to create and maintain the records. Like the Data Archaeology Strategy, Emulation presumes the on-going readability of electronic documents, which means periodically transferring electronic records to new storage for as far into the future as necessary.

The Digital Archaeology and Emulation Strategies both share the objective of preserving "digital executable originals." They differ primarily when reverse engineering or emulation will occur. Emulation as, Rothenburg defines it, is intuitively attractive but it currently has little market place support. (23) Although the library and archives communities have shown considerable interest in Emulation, (24) it fundamentally is an untested "theoretical approach" that will require the development of new tools and approaches. The success of Emulation as a long-term archival preservation strategy depends upon the results of empirical studies currently under way and the interest and support of mainstream technologies to integrate it into other tools that serve business needs and requirements. Until this happens, there is little to be gained by the SIA adopting the Emulation Strategy.

2.3.3 Migration

The third strategy, known generally as Migration,(25) is less concerned with static "executable digital

originals" and focuses instead upon maintaining dynamic electronic records that can be rendered and processed with whatever technologies current and future market places support. A key component in the Migration Strategy is the notion that more robust and efficient vendor technology neutral standards that support backward compatibility across technology generations will replace older standards. An example of this is XHTML, which is likely to replace HTML. At some point in the future XHTML will be replaced by a newer vendor technology neutral standard that is more robust and efficient. Backward compatibility within vendor technology neutral formats will make it relatively easy to migrate Web XHTML pages to the new format. This exercise, of course, must be repeated indefinitely into the future. In this regard, the Migration Strategy acknowledges that some loss of software functionality associated with electronic records as well as some diminution of the "look and feel" of the electronic records may occur over successive migration generations. Therefore, the Migration Strategy requires documentation of any diminution of the "look and feel" of electronic records or the loss of software functionality in a preservation history audit trail. Like the Data Archaeology and Emulation Strategies, the Migration Strategy also requires ensuring the on-going readability of electronic records by periodically transferring them to new storage media.

The preservation of "executable digital original" electronic records, which is reflected in the Data Archaeology and Emulation Strategies, is intuitively attractive but not very practical in today's world. The Data Archaeology Strategy can be employed as a short-term strategy (five to ten years) but is not viable as a long-term strategy. The Emulation Strategy is still in the theoretical stage and requires substantial software development before it can be a viable preservation strategy. The only viable long-term preservation or archiving strategy today is Migration. (26)

---

3. SMITHSONIAN INSTITUTION WEB RESOURCES

3.1 Overview of Smithsonian Web Sites

Background. The SI public Web site (http://www.si.edu) consists of an enterprise-wide home page from which viewers can navigate to more than seventy-five separate public Web subsites. These seventy-five Web subsites contain thousands of HTML pages that occupy between thirteen (13) and twenty (20) GigaBytes of storage. (27) The most popular Web sites are:

- The National Air and Space Museum (http://www.nasm.si.edu),
- The Smithsonian Institution Magazine (http://www.smithsonianmag.si.edu),
- The Smithsonian Astrophysical Observatory (http://sao-www.harvard.edu/sao-home.html),
- The National Museum of Natural History (http://www.mnh.si.edu),
- The National Zoo (http://natzoo.si.edu),
- The National Portrait Gallery (http://www.npg.si.edu),
- The National Museum of American History (http://americanhistory.si.edu), and

- The Office of Education (**http://scemsweb.si.edu**). (28)

Structure. The SI public Web site structure consists of four separate but interrelated means for supporting public Web sites. (29) First, several museums support their own Web server and no Web material is hosted on the central SI public Web server. Instead, there is a navigation or "transportation" link from it to the specific museum server. The National Air and Space Museum is an example of this relationship. Second, some museums and research centers have both their own Web server and host information on the central SI public Web server. The exact relationship varies considerably, ranging from a full mirror site, partial mirroring/partial split data, and a general split of unique information on each server that links back and forth. The National Zoo and the National Museum of Natural History mirror some of their Web site material with the central SI public Web server. The third category includes some museums and bureaus that do not have a Web server so each entire site is hosted on the central SI public Web server. The Hirshhorn Gallery and the Smithsonian Library are examples of this relationship. Fourth, two SI research centers - the Smithsonian Astrophysical Laboratory and the Smithsonian Tropic Research Institute - are physically located on servers that are external to the central SI public Web server.

Management. The public SI Web environment consists of the central public Web server under the technical management of the SI Webmaster and individual museum, department, and research center Web sites (either hosted in their entirety, partially, or not at all on the central SI public Web server) under the management of their respective webmasters. Technical management includes network support, acquisition of new software, system upgrades, and the like. Each museum, department, and research center may have space on the central SI public server and individual webmasters have total control over the content of their respective sites (subject, of course, to policy and procedures in place) and frequency of updates. Most individual webmasters also manage a departmental server that is used for developmental purposes and testing of material before it "goes public." Also, as noted above, some museums and research centers maintain a public Web server under the management of the respective webmasters that can be accessed through a link on the SI central public server page.

3.2 SI Web Policy

In May of 1996 the Office of Information Technology issued a document entitled "Web Development and Operational Guidelines." (30) The guidelines cover a number of issues including the following:

- The Office of Information Technology manages the "Smithsonian Management-level home page" to ensure a "common look-and-feel" across the institution,
- The creation, maintenance, and enhancement of HTML pages is the responsibility of appropriate "Research Center/Bureau/Major Office/Museum,"
- Webmasters are responsible for the accuracy and content of their respective HTML pages, and
- Bureaus, museums, and major offices are encouraged to purchase or lease their own Web servers.

More specific guidance HTML page standards specify that:

- Each page should include the date it was last changed and whom to contact for information,
- HTML code that is vendor specific shall not be used, and
- All links within the same server should be relative (not absolute).

The guidance document makes no provision for the creation of back-ups or the archival preservation of Web-based material.

Four museum/research center Web sites have a written Web policy. The National Air and Space Museum has "General Guidelines for Web Sites" that largely focus on the responsibilities of a "Web Team" to "foster participation by the divisions in the NASM web site and to establish standards for the development of content." (31) The guidelines make no provision for archival preservation. The "National Museum of Natural History Internet Publishing Policies and Procedures" is quite extensive (more than 20 pages long) and covers a number of procedural matters. For the purpose of this study, the most important section is called "Server Administration and Maintenance." One provision in this section calls for the regular copying of "software and data files to tape, so they can be restored in the event of a hardware failure or data corruption problem." (32) In addition, major revisions to the NMNH Web Home Page and "high-level navigation and informational pages shall be archived to tape for future review or use." Probably the most significant procedure declares that information providers may request that a 'snapshot' archive be created of a particular presentation. This archive service captures the presentation as it exists on the National Museum of Natural History Web at a specific point in time. The archive is created in Unix "Tar" format and therefore can be directly reloaded on the Internet server, or any similar Unix server. The archive media (currently, tape) can be kept by the information provider of files in the Webmaster's archive library. (33)

The Web Design Policy for the Smithsonian Environmental Research Center includes a requirement that "Information such as the tools used to create the HTML, sources of inline multimedia elements and nonstandard formatting should also be documented within the page itself" and that links to other pages in the SERC Web site "should be coded with relational (sic) addresses rather than with absolute addresses." (34)

The Web Development and Operation Site Guidelines of the Smithsonian Institution Libraries (SIL) is a carefully thought out document that covers a number of topics including accuracy of the content, mark-up guidelines for SIL on-line publications, and links to best sources. Among the more noteworthy provisions are:

- HTML markup will follow basic guidelines so as to be XML compliant,
- All WWW documents are signed with the initials of the author, compiler, or editor with the date of publication or update,
- The compiler is responsible for maintaining a record of all changes to the Home Pages for which s/he is responsible, and
- "Backup" and archiving functions should be conducted to ensure continuing availability and

access to WWW materials as standards and technology evolve. (35)

Unfortunately, "archiving functions" are not described so it is unclear exactly what these functions entail.

Aside from these four Web sites and the central SI public Web server, it is not known what formal policy and guidelines, if any, are in effect for the remaining SI public Web sites. Certainly, it would be prudent for these Web sites to adopt a common policy and guidelines that could include:

- All HTML pages should identify the author/creator and date of last change,
- Vendor specific HTML code shall not be used,
- All links within a server shall be relative, not absolute,
- HTML markup shall be XML compliant,
- Archival copies of Web sites shall be periodically created and maintained, and
- Major revisions to a Web site should be fully documented.

3.3 National Air and Space Museum (NASM)

The SI central public Web server does not host the NASM Web site but has a "transportation link" that takes users directly to it. Once in the Web site, viewers have a number of options, including a new online exhibition, "Looking at Earth," visitor information, a description of current exhibitions in the gallery and online exhibitions, collections and research, and educational services among others. Visitor information includes a calendar of events prepared by the Visitor Information and Associates' Reception Center and a calendar of events prepared by the NASM staff. There is considerable duplication in the two calendars (e.g., both calendars show the lecture on April 21, 2001 of Retired General Paul W. Tibbets, Jr.) but the NASM prepared calendar of events includes events that are scheduled after the calendar event deadline. The Garber Restoration Facility features "Live Web Cam" that streams real-time images of restoration work to the viewer's monitor for display.

Perhaps the richest elements of the NASM Web site are the Search and Features section displayed toward the bottom right side of the page. (36) The search feature includes a keyword search function that uses the Verity search engine. The index itself is created with Verity and is under the total control of the NASM. The Features section includes a listing of Online features, some of which replicate closed exhibitions or current exhibitions in the Exhibition Galleries. Equally as important is a listing of Special Online Resources from the National Air and Space Museum that is composed of History and Reference Resources, Image Collections, Educational Activities, Publications, and Links to More Information. The History and Reference Resources include African Americans in Aviation History, Women in Aviation and Space History, National Air and Space Museum - Archives Division (documentary materials), and Oral History on Space, Science, and Technology. A catalog description of the program along with the text of selected oral interviews can be viewed online. The Educational Activities component features several "online galleries" for classroom and home use. In addition there is a link to a third-party Web site called Science Education Gateway that is a collaboration between science museums, researchers, and educators.

The publications program of the NASM includes museum publications (e.g., Smithsonian History of Aviation and Spaceflight Series) and links to the Home Page of the Air and Space Magazine and the Smithsonian Magazine. Rounding out the Online Resources is a series of links to other Internet sites that provide information on aviation and space related topics.

Certain access and preservation issues associated with the NASM Web site merit brief discussion. First, there is an inconsistency in providing revision information at the bottom of HTML pages. In some instances there is no revision information. In the vast majority of instances the date of last revision along with the author of the revision (usually initials) is provided. Second, Vicki Portway, the NASM Webmaster, reports that the active NASM Web site currently is between 3.7 and 4.0 GB and that she backs up the NASM Web site each month on CD-ROM and an 8-mm tape. She also maintains a history log of all changes made to the site. Third, ColdFusion is the authoring tool used for NASM created HTML pages. Most of the HTML pages, including those constructed by contractors, are in version 4.0 so it appears that there are no legacy HTML pages that require conversion. Fourth, except for "Visitor Comments," virtually all of the current NASM HTML pages are static pages.

3.4 National Museum of American History (NMAH) (37)

As noted earlier, the central SI public Web server hosts the NMAH Web site. The NMAH home page consists of a list of theme areas (e.g., "Virtual Exhibitions" in which there are twenty-four separate ones), "You and the Museum" information, Collection, Scholarship & Research, Search Site, and two Exhibitions. The "You and the Museum" information includes a NMAH calendar of events that lists more events (like the NASM) than does the NMAH Calendar of Events on the SI Home Page. The core of the NMAH Web site is under "Collections, Scholarship & Research," which actually consists of three Web sub-sites - the Archives Center, the Lemelson Center, and Curatorial Divisions - hosted on three separate servers, each of which has its own Webmaster. Virtually all of the HTML pages of these three subsites and various functionalities associated with them are static pages that display the same information to all viewers. The one instance of database integration is the Vidal Virtual Exhibit and it runs under Filemaker Pro Database, a proprietary software package. (38) In addition, several NMAH Web sites now capture visitor details and support on-line queries for research information.

The total number of HTML pages that comprise these three Web components of the National Museum of American History public Web site is approximately 10,000 with almost one-half of them housed on the Curatorial Divisions Web server. It is not known if any of these 4600 HTML pages were authored with HTML 2.0 or HTML 3.2. In fact, there is very little that can be confirmed about which version of HTML was used because most of the design and construction of HTML pages is done by outside contractors.

The NMAH Web Home Page features a search function linked to a third-party service called Master. com that "crawls" the NMAH site weekly to create an index to all of the Web site material. All of the index data with their related links are maintained by Master.com at its site. When a viewer enters a search term and executes the Search function, the query is transparently moved to the Master.com site

where the search is conducted and a list of "hits" along with their NMAH links is retrieved but only the "hits" are displayed. Clicking on a "hit" will take a viewer to the appropriate HTML page. This is convenient for both the NMAH and viewers but it poses a problem for archival preservation because the content of the Master.com Web source changes weekly and the database is proprietary.

Documenting changes in HTML pages varies somewhat. Over the last year or so the Archives Center Webmaster and the Lemelson Center Webmaster have begun keeping a log of changes to HTML pages and the Web sites. The Archives Center HTML pages consistently identify the date of last revision and the author of the change while the HTML pages of the Curatorial Divisions and the Lemelson Center are somewhat inconsistent in this regard. Over the last two years the Archives Center Webmaster has created a monthly copy of the entire site on CD-ROM. The Lemelson Center Webmaster began making monthly copies about six months ago. From time to time the Curatorial Division Webmaster has made another copy of the Web site on CD-ROM.

3.5 The Freer Gallery of Art and the Arthur M. Sackler Gallery (ASIA)

The central SI public Web server hosts the ASIA Web site in its entirety. When viewers enter the ASIA Home Page they have a number of options, including Exhibitions, Calendar, Ask Us, Education, Gallery Shop, and Cherry Blossoms. The Exhibitions include a listing of all the gallery exhibitions as well as three on-line exhibitions (Devi, Puja, and Preserving Ancient Statues from Jordan.) One of the gallery exhibitions, India: Through the Lens of Photography, 1840 - 1911 includes a feature called "We want to see India through YOUR Lens" that invites viewers to submit a digital version of their own photographs along with descriptive text for posting on the Web site. The Calendar is maintained by the ASIA staff and includes events and details not covered in the calendar prepared by Visitors and Associates Reception Center.

The ASIA Web Home Page features a search function linked to a third-party service called Atomz.com that "crawls" the entire site periodically to create an index to all of the Web site material. Like NMAH's use of Master.com, all of the index data with their related links are hosted on the Atomz.com server. When a viewer enters a search term and executes the Search function the query is transparently moved to Atomz.com where the search is conducted and a list of "hits" along with their ASIA links are retrieved but only the "hits" are displayed. Clicking on a "hit" will take a viewer to the appropriate HTML page. This is convenient for both the ASIA and viewers but it poses a problem for archival preservation because the content of the Atomz.com Web source changes weekly and the database is proprietary.

Certain access and preservation issues associated with all SI Web sites merit brief discussion. The date of last revision but not the author's name is posted at the bottom of each HTML page. The ASIA Web site consists of about 300 static HTML pages and one dynamic page that are backed up every two months on one CD-ROM. These HTML pages are written in HTML 4.0. The design and construction of HTML pages is done internally by ASIA staff. The primary Web authoring tool is ColdFusion.

---

4. SELECTION OF SI WEB SOURCES FOR ARCHIVAL RETENTION

4.1 Web Sources as Records

A term encountered frequently in the Internet world is "web publication," which generally refers to any kind of digitally encoded information that is created and made available to the public through a communications network, (39) which clearly could include records. However, traditional archival theory and practice have distinguished between publications and records. (40) The rationale for this distinction between publications and records is that the former typically involve multiple copies of information, no one of which is unique, that are not produced in the ordinary course of business, and, therefore, do not rise to the level of a record as evidence.

In a Web environment this distinction is arbitrary and serves no useful purpose because it confuses the means of publications with the purpose and content of publications. The means for "Web publications" require use of the Internet that by its very nature virtually guarantees many people will see such publications and they may end up listed on multiple Web sites. But this is incidental both to the purpose of the publication as well as its content. If the purpose of Web publication is to convey information in accordance with an informal or institutional policy, or the content of the publication clearly is evidence of an institutional program or function, then it does rise to the level of a record. Thus, for the purposes of this study a "Web publication" may indeed be an electronic record. It must be emphasized that the record status of a "Web publication" applies only to those Web sites under the direct control and management of the institution. Third-party Web sites linked to an institutional URL/Home Page are not under the control or management of the host server and therefore would not be records of the host institution.

Some archivists hold the view that even when a "Web publication" rises to the level of records, if the material is captured in an electronic recordkeeping system or is duplicated in paper-based records, (41) then it is redundant and need not be preserved. This view is wrong for the following reasons. First, it overlooks a fundamental feature of Web-based records: the Web context within which electronic records are presented. Second, this context includes hyperlinks that are part of the context of the use of electronic records in an environment that cannot be duplicated in an electronic recordkeeping system. The juxtaposition of text and a related image on the same HTML page (interlined or a hyperlink) is a contextual linkage, that is, a "look and feel," not likely to exist in either the text or the image in an electronic recordkeeping system. Third, the Web-based text that is duplicated in a word processing document stored in an electronic recordkeeping system will have the same content but the rendering presentation of the two can differ substantially, depending upon the complexity of the document structure and the Web authoring tool used. (42) Therefore, this is a difference of the "look and feel" of Web based records vis-à-vis their counterparts in an electronic recordkeeping system. From an archival science perspective, it can be argued that Web-based text or images that duplicate information in an electronic recordkeeping system in fact are new records because their presentation, structure, and context will be different. (43)

In those instances where the content of HTML pages is generated on the fly by a database or Active Server Page there should be no debate about the record status of such "Web publications." Nor for that matter should there be any question about the record status of "Web publications" (e.g., Electronic Exhibits) that are "born digitally" and remain so throughout their retention.

This study argues that the Web context in which SI Public Web sites functions raises the level of Web site electronic material to that of an electronic record. Some SI HTML pages may be trivial without any long-term consequences and therefore may be retained for only a short period of time. Other SI HTML pages will have greater institutional significance and therefore merit designation as "permanent records."

4.2 Smithsonian Institution Archives Appraisal Methodology

In September 1999 the SIA issued a report on **Appraisal Criteria and Methodology** that is based on a functional analysis approach. The report identifies four broad functions of the SI:

- Assuring institutional continuity,
- Acquiring and maintaining the national collection,
- Conducting and supporting original research, and
- Diffusing knowledge.

In recent years, many archivists have promoted archival appraisal that is based upon a functional analysis of an organization but few archivists have actually completed a sound and coherent functional analysis that can drive archival appraisal. The SIA is one of a handful of archives to actually frame archival appraisal within the context of a functional analysis. Within this context, SIA archival appraisal decisions are based on "the overall value of records as documentary evidence of the functions and activities of the office, …" (44) The report identifies a number of activities subsumed under the four functions of the SI that produce records that are likely to be of "enduring value." For example, the function of diffusing knowledge is subdivided into three activities:

- Exhibitions: Addresses activities concerning conceptualization, planning, funding, execution and evaluation of exhibitions.
- Productions: Addresses activities concerning conceptualization, planning, and funding of productions, including print, electronic, and broadcast media. SIA will attempt to collect designated series of core productions, and others, as opportunity permits.
- Events: Addresses activities concerning conceptualization, planning, funding, and execution of events, including educational courses and tours, symposia and seminars, festivals, lectures, and performances.

Although the report refers to the appraisal of electronic records, (45) the guidance is general and does not explicitly delineate the archival appraisal of SI Web sites and pages. (46) For example, under the function of "Diffusing Knowledge" certain records associated with exhibitions are considered to be of

"enduring value" but the language implies paper-based documentation (e.g., "Exhibition Catalogs," "Exhibition floor plans," and "Visitor Comments"). (47) Nonetheless, SI Web sites do engage in "Diffusing Knowledge," especially through exhibitions, and the "enduring value" of SI public Web sites and HTML pages can be extrapolated from those SI records already identified to be of "enduring value."

Archival appraisal and retention of records usually is done at what is called the record series level. A record series is a body of documentary material that arises out of activities associated with a clearly defined function or process and therefore may share a commonality in terms of how the records are filed, the subject matter of the records (e.g., contracts or outgoing correspondence), and the like. This implies, among other things, that the content, structure, and context of a record series are under the administrative management and control of an individual(s) who is accountable for the accuracy of the content and presentation of the records. In the SI context a public Web site and its associated HTML pages are the means for the diffusion of knowledge, one of the four functions that the SI performs, and therefore they rise to the level of an archival record series. The mapping of the NASM Web site and associated HTML pages to SIA appraisal criteria that follows in the next section is rooted in the notion that the NASM Web site is a record series that is the by-product of the function of the diffusion of knowledge.

4.3 Application of SIA Appraisal Criteria to National Air and Space Museum

Defining the NASM Web site and HTML pages as equivalent to a records series excludes from consideration all Web-based material that is not under the administrative control and management of the NASM Webmaster. Accordingly, the following Web sites and HTML pages that are external to the NASM Web site are not considered part of the NASM Web record series:

- Science Education Gateway,
- Air and Space Magazine,
- The Smithsonian Magazine,
- Links to More Information,
- Educational Links (Online Activities),
- Hot Links to Aerospace Web sites (HTF Resource Center),
- SI Calendar of Events (Visitor Information and Associates' Reception Center),
- History of Aviation Series Publications (Publications), and
- Solar System Series Publications (Publications).

The exclusion of these Web sites and associated HTML pages from the NASM Web record series, however, does not mean the existence of their linkage is of no consequence. In fact, these links constitute "gateways" to a broader context of the NASM Web record series and knowledge of their existence should be documented in metadata. As a practical matter this can be done by changing the URL for each third-party Web site to a NASM HTML page that briefly describes the link and its general content and gives the URL that was operational at the time of transfer to the archives Web server.

Undoubtedly, some of the material on the NASM Web site duplicates information held in various offices and departments of the NASM. For example, Frequently Asked Questions (FAQ), the Albert Einstein Planetarium Schedule, National Air and Space Museum Fact Sheet, Museum Maps, and Special Events, among others are records created and maintained by other units of the NASM. Under the traditional concept of the "Office of Origin" or "Office of Primary Responsibility" Web postings that duplicate records held in other NASM offices do not have enduring value and therefore can be disposed of. The information content of records created and used by a NASM unit may be the same as that of Web postings but the context of use is substantially different, which gives rise to a record status of enduring value.

All of the remaining Web material on the NASM Web site comprises a Web record series and should be retained for its enduring value. The history log of changes maintained by the NASM Webmaster and comments and feedback received from visitors is part of the NASM Web record series and should be retained. The monthly Web site copies (CD and 8-mm tape) the NASM Webmaster maintains have short-term operational value and need not be retained as part of the NASM Web record series.

The Garber Restoration Facility HTML page, which is part of the NASM Web record series, includes a live, real-time video of the facility although individual viewers may have only fifteen seconds (15) of viewing time. Capturing the continuous live feed of the Garber Restoration Facility would create very difficult storage problems and technical issues. One possible solution would be to capture one instance of the live feed (15 seconds or so) at the time a snapshot is taken of the Web site.

5. RECOMMENDATIONS

This chapter of the report delineates six areas of recommendations for follow-on activity that reflect a life cycle management approach. The six areas are:

- A SI wide Web policy for the life cycle management of its Web sites and HTML pages,
- Archival considerations in the design and authoring of Web sites and HTML pages,
- Appraisal considerations for the selection of SI Web sites and HTML pages for archival preservation,
- Archival capture of SI Web sites and HTML as records,
- A SI Web site and HTML archival preservation strategy, and
- SI Archival Web site and HTML principles and best practices

Recommendation areas 2, 3, 4, 5, and 6 presume that a policy framework will be established that supports specific recommendations or there is a consensus that they are worth doing.

5.1 SI Archival Web Policy

The SI has a general policy regarding the design, maintenance and use of Web sites and HTML pages but virtually nothing regarding the archival preservation of Web sites, HTML pages, and associated documentation (metadata). Developing a SI policy document that links the design and use of SI Web sites and HTML pages with their archival preservation should correct this deficiency. The formulation of a comprehensive life cycle management policy governing SI units (museums, research centers, and other organizations) that create, use and maintain Web sites, HTML pages, and associated metadata and other records is beyond the scope of this study. Nonetheless, it is possible to identify areas that such a policy document should address.

Scope. This policy should apply to all SI entities that maintain a public Web site hosted on either the central SI public Web server or on an internal independent public Web server and use the site to provide services, products, and information to the public. This would exclude those Web sites whose content is not controlled by a Webmaster who is accountable to the SI.

Objective. This policy should articulate clearly that it is the intent to ensure that trustworthy and usable documentation of Web-based activity is created and retained as part of its corporate memory and thereby meet legal and institution obligations and community expectations for access for as long as may be required.

Special Features. As noted earlier, the delineation of a comprehensive SI policy for the archival preservation of Web-based records is beyond the scope of this study. Nonetheless, several special features can be identified that this policy should take into account. They include:

- The life cycle management of SI Web-based records in which a Web site is defined as a record series,
- Use of the Dublin Core or its functional equivalent to support uniform retrieval of HTML pages across all SI Web sites,
- Archival appraisal of SI Web sites and HTML pages that incorporates functional analysis,
- Archival capture of SI Web sites and HTML pages on an archives Web server,
- Protection of the integrity of SI Web sites and HTML pages over time despite changes in technology platforms,
- Minimization of software dependence through the use of vendor technology neutral formats,
- Minimization of hardware dependence through periodic transfer of Web-based records to new storage media, and
- Capture of audit trail metadata about preservation activities taken to extend the usability of Web-based records.
- Use the Dublin Core Metadata (or its functional equivalent) for retrieval/cataloging purposes,
- Minimize software dependence through the use of vendor technology neutral file formats
- Minimize hardware dependence through the transfer of electronic records to new storage media for which there is a strong market presence, and
- Capture of audit trail metadata about preservation activities taken to extend the usability of electronic records.

5.2 Long-term Access Considerations in the Design and Authoring of SI Web Sites and HTML Pages

The Life Cycle Management of usable and trustworthy SI Web sites and HTML pages requires that long-term access considerations be taken into account in the design, authoring, and maintenance of SI Web sites and HTML pages. These long-term access considerations include the definition of a SI Web site as a record series, the utilization of technology independent standards and standardized metadata (e.g., Dublin Core Metadata), and periodic creation of copies. The failure to incorporate long-term access considerations in the design, authoring, and maintenance of usable and trustworthy SI Web sites and HTML pages will make their preservation more difficult and costly. Implementation of the following recommendations will minimize this difficulty and expense.

Recommendations:

- Use the Dublin Core Metadata (or its functional equivalent) for descriptive/retrieval purposes,
- HTML markup should be XML compliant,
- Avoid the use of a third-party proprietary search engine that is not under the control of a SI Webmaster,
- All links within the same Web site should be relative (not absolute),
- Copies of Web sites should be periodically created and maintained, and
- Major revisions to a Web site should be fully documented.

5.3 SI Archival Appraisal of Web Sites

Archival appraisal of SI Web sites is important because it is the means for identifying those Web sources that are essential to protecting the SI corporate memory and meeting legal and institutional obligations and community expectations. The archival principles underlying the SI Appraisal Methodology include:

- A record series is the fundamental unit of appraisal,
- A record series consists of records that are the by-product of activities that occur in carrying out one of the four functions of the SI,
- A records series consist of records that are under the control of individuals who are accountable to an organizational component of the SI, that is they are created or received and maintained by individuals acting in their official capacity in carrying out one of the four functions of the SI,
- A records series is "owned" by the organizational unit with the primary responsibility for its creation and management and should not be intermingled with other records series, and
- A records series may contain records duplicated in another records series but their new context of use, presentation, or delivery makes them new records that are incorporated in this records series.

Based upon these archival principles, this study argues that a SI Web site and its associated HTML pages is a SI records series when it is under the control and management of a SI employee acting in his/her official capacity in carrying out one of the four functions of SI. The SI organizational unit (or its

surrogate) with the responsibility for creating and maintaining the content of a Web site and its associated HTML pages "owns" the Web record series. In the Smithsonian Web environment there are as many Web record series as there are Web sites and associated HTML pages whose content is created and maintained by a SI organizational unit. For example, the Smithsonian Home Page is one records series with navigation paths to virtually all of the Web records series that comprise the SI Web environment. The Smithsonian Home Page is under the control of the SI Webmaster while each museum or research center that is linked to the Home Page is under the control and management of a Webmaster. Links from a specific SI Web records series to another SI Web records series or to Web sites outside the SI, are pointers to supplemental or complementary information that is not part of that specific Web record series. A SI Web records series, therefore, is an organic whole rather than an aggregation of HTML pages. Web sites and associated HTML pages should be considered a SI Web records series under the following conditions:

- Each SI Web site with a unique URL is a Web records series,
- Each SI Web records series consists of a Web site and associated HTML pages whose content is created and maintained (i.e., "owned") by an identifiable SI organizational entity, and
- Any Web records series or third-party Web sites and HTML pages with links to a specific SI Web site and its associated HTML pages are not part of that specific SI Web record series. (48)

There is an exception to the third bullet above when a Web plug-in that is essential to the overall functionality effectiveness of a Web site or HTML pages is downloaded from a vendor on demand. Without a functional Web plug-in (e.g., Shockwave) then part of the look and feel of a Web site or HTML pages will be diminished. (49)

Recommendations:

- Adopt the definition of a SI Web records series that has a unique URL whose content is created and maintained (i.e., "owned") by an identifiable SI organizational entity,
- Generate a SI Web site map and identify each SI Web records series,
- Generate a Web site map for each SI Web records series and identify the HTML pages that belong to each records series, and
- Schedule each SI Web records series for permanent retention and transfer to the SIA Web archives repository.

5.4 SI Archival Capture of Web Sites

There are two strategic options for the archival capture of Web sites and HTML pages that can be characterized as content driven and event driven strategies. (50) The content driven strategy emphasizes the capture at specific point in time of the content and any functionality of Web sites and HTML pages that are required for their subsequent presentation to viewers. The content driven strategic option is appropriate where the goal of archival preservation is to ensure that future users can view Web sites and HTML pages as they existed at a particular point in time in the past. The content driven strategy works

best with static Web sites and HTML pages whose content changes infrequently. In contrast, the event driven strategy focuses upon transactions between users of dynamic Web sites and HTML pages that involve some level of interaction between users and a Web site or pages. A case in point is an Active Server Page application or interactive database that produces customized HTML pages based upon transaction parameters that include business rules, user access rights, and the specific view of the database required (derived tables) for the transaction. Each instance of such transactions, therefore, represents a unique user view of the Web site and HTML pages that physically exists only at the time of the transaction and its transmission to the viewer. The event driven strategy works best with dynamic, interactive Web sites and HTML pages.

The implementation requirements for these two capture strategies differ substantially. The primary requirement for the content driven strategy is to capture a complete and accurate copy of all of the content and associated application functionality of a Web site and HTML pages so that they are frozen in time. The primary requirement for the event driven strategy is to capture the transaction parameters - the query, the business rules, user access rights, the specific "view" state of the database at the time of the transaction - and the underlying application so that the transactions can be reconstructed on demand in the future.

These two strategic options specify what should be captured, not when the capture should occur. The time of capture depends upon the scope and frequency of changes to Web sites and HTML pages. With static Web and HTML pages substantive change is likely to be infrequent so an initial base line capture would be sufficient until there is a major redesign of the site and pages, at which time another "historical" snapshot would be necessary. Between the initial base line capture and the redesign of the site and pages minor changes, such as monthly event calendars, can be captured and written to a history log change file. The "historical" snapshot and history log change file should provide adequate documentation of the content of static Web sites and HTML pages.

The capture of dynamic Web sites and HTML pages can be done in two different ways. The first involves capturing "on the fly" each set of transaction parameters, the derived tables (database "view"), and the database application software. The second way involves capturing "on the fly" each set of transaction parameters and the entire database at specified intervals (e.g., daily) along with the database application software. A daily capture of the database presumes that it is updated at regular intervals, say, every day at 12:00 AM. Based upon today's technologies the infrastructure required for the capture of dynamic Web sites and HTML pages is likely to be substantial. Also the storage costs for preserving executable Web sites and HTML pages over time are likely to be equally substantial.

The issue of capturing dynamic SI Web sites and HTML pages is of little consequence because currently virtually all of the SI Web sites and HTML pages (on the order of 95% or higher) are static. Over the next five years or so this may change as more dynamic Web sites and HTML pages are created but this is not an issue that needs to be addressed now.

Recommendations:

- Capture a snapshot of each SI Web site and associated HTML pages as a base-line,
- Establish a history change log for each SI Web site and document all revisions,
- Capture a new snapshot of any new HTML pages added to a Web site, and
- Capture a new snapshot of a Web site when it is redesigned or there are major changes.

5.5 SI Archival Preservation Strategy

As noted earlier (Chapter 2) there are only two viable preservation strategies that the SIA may initiate. The first, the Data Archaeology Strategy, would entail the transfer of a copy of SI Web sites and pages of archival value from the public SI Web server or museum/research center/departmental server to a dedicated electronic archives repository. The SIA would control access to the contents of the archives to protect the integrity of the material and collect documentation about the operating systems and application software along with information about vendor HTML extensions and plug-ins used. The preservation activity the SIA would undertake would be to:

- Collect documentation about each Web site and its associated HTML pages,
- Generate an ASCII listing of all HTML pages (site map) and their URLS,
- Identify the Web authoring tool(s) used to create the material,
- Identify the version of HTML used,
- Identify any proprietary plug-ins,
- Store each Web site and associated HTML pages on an archives repository Web server,
- Create a second copy of the Web sites and HTML pages and transfer to a storage medium such as Digital Linear Tape,
- Monitor the readability of Web sites and HTML pages on the electronic archives repository and copies on a removable storage medium such as Digital Linear Tape, and
- Transfer Web sites and HTML on the current archives repository server to a new server at some point in the future and transfer the copies to new storage media at some point in the future.

The Migration Strategy, the second viable preservation strategy, would involve at least four separate but closely related activities. The first would be for the SIA to adopt HTML 4.02 as its archival storage format and to encourage all SI Webmasters to update all of their active Web sites and pages to HTML 4.02. Second, when XHTML becomes available the SIA should adopt it as its archival storage format and encourage all SI Webmasters to update all of their active Web sites and pages to XHTML. Third, when a copy of these Web sites and HTML pages is transferred to the SIA Electronic archives repository, the SIA would assume responsibility for ensuring that the sites and pages are converted to new vendor technology neutral formats that replace XHTML. In this regard, it should be noted that migration is a process that will need to be repeated as far in the future as is necessary. The fourth activity would be for the SIA to retain at least one back-up copy of all Web sites and HTML pages on removable storage media (e.g., magnetic tape) in a Tar format that will protect the directory structure.

Recommendations:

- Adopt the Digital Archaeology Strategy as a means of acquiring immediate custody of SI Web sites, HTML pages, and associated metadata (and documentation),
- Transfer a SI Web records series that is not overly complex to an electronic archives repository so that SIA staff may obtain hands-on experience, and
- Implement the Migration Strategy over the next two to four years as resources are made available.

5.6 SI Preservation Principles and Best Practices

The following section reviews nine preservation principles and recommends a number of best archival preservation practices.

5.6.1 Establish an electronic archives repository

It is axiomatic that electronic records should be transferred from an operational environment to an electronic archives repository. Electronic records stored in an operational environment are still under the control of their creators and users and are subject to alterations and destruction either by design or accident. In contrast, an electronic archives repository is a "trusted third party" whose mandate is to ensure the preservation of trustworthy and usable electronic records as evidence for as long as necessary. Typically, this involves "read only" access to digital material in the repository and protection of the physical security of the repository. A base line requirement for the preservation of trustworthy and usable Smithsonian Web Sites and HTML pages is the creation of an SIA Electronic Archives Repository. Such a repository may be managed and staffed by the SIA or it could be outsourced. A risk analysis of these two archival repository alternatives should be conducted to determine which is more cost-effective.

If the SIA manages and staffs the electronic archives repository then some attention should be given to where the repository fits into the Smithsonian Institution Technical Reference Model Strategy. (51) Under this model it appears that an electronic archives repository would fall under Data Management Services, which includes Data & Records Management, Data Repository, Data Quality, Data Warehousing, and Database Management, all of which appear to be applications in an operational environments. An electronic archives repository should not be part of an operational application environment. Of course, a firewall could be put in place that would have the practical benefit of removing the Electronic Archives Repository from an operational application environment. Nonetheless, the preferred solution would be to add an Electronic Recordkeeping Services along with the appropriate standards to the Technical Reference Model. This would give greater visibility to the preservation of trustworthy and usable SI Web sites and HTML pages and to the preservation of trustworthy and usable electronic records as evidence in general.

Recommended Best Practice:

- Establish an Electronic Archives Repository under the control of the Smithsonian Institution Archives (SIA),

- Conduct a cost benefit analysis to determine if the repository should be staffed and managed directly by the SIA or outsource to a vendor, and
- Add an Electronic Recordkeeping Services to the Smithsonian Institution Technical Reference Model.

5.6.2 Transfer Web sites and HTML pages to the SIA Archives Repository

The transfer of SI Web sites and HTML pages from an operational Web server environment to an electronic archives repository should be early in the records life cycle. All of the contextual metadata (i. e., documentation) that supports an understanding of the status of the records during their creation, use, and transmission should be included in the transfer. This is particularly critical when Web sites and HTML pages were authored using non-standard HTML tag extensions, unique screen layout features, and Web plug-ins that no longer are executable. It is important, therefore, to document why full reconstructability is not possible so that future users will have confidence in the trustworthiness of the Web sites and HTML pages they access.

The documentation that should be compiled extends beyond these particular features. Appendix 1 in this study lists a series of questions about the design, use, and maintenance of Web sites and HTML pages that should be incorporated into requests for documentation. Much of this information should be included in the metadata that documents the context of the transferred Web sites and HTML pages.

The preferred means for transferring Web sites and HTML pages to the custody of the SIA archives repository is the File Transfer Protocol, which already is used to transfer Web sites and HTML pages from a development server to the central Public Web server. It would be a relatively straightforward process to establish an archives Web server and transfer to it fully functional Web sites and HTML pages that are hosted on the central SI public Web server or on standalone servers maintained by departments/museums/or research centers. In the first instance the SI Webmaster could do this transfer while in the second instance individual Webmasters would have to do the transfer. In either case, use of the FTP would require little effort on the part of anyone. The major difficulty is likely to be encountered when partial Web sites from the same organizational entity are hosted on the central SI public Web server and a departmental standalone server. Combining these two components into a fully reconstructed Web site would be the responsibility of the SIA.

An alternative, albeit much less efficient, would be to write a Web site with all its associated HTML pages to removable storage media (optical or magnetic) in the TAR format. Use of the TAR format ensures that the file directory structure of the Web site is transferred intact. The chief advantage of writing a Web site and associated HTML pages to removable storage media is that it is not necessary to know what the Target system is at the time of writing. The Web site and associated HTML pages can remain in the Tar format indefinitely, subject only to periodic media renewal, and migrated to any technology platform that can import data in the TAR format.

Recommended Best Practice:

- Transfer Web sites and HTML pages to an electronic archives repository as early into the life cycle as possible,
- Use the FTP whenever possible to transfer functional Web sites and associated HTML pages from an operational Web server to the electronic archives repository,
- Write functional Web sites and associated HTML pages to removable media in the Tar format when the target technology platform is not known,
- Require that appropriate documentation accompany the transfer of the Web sites and HTML pages, and
- Utilize the questions in Appendix 1 in specifying what documentation should be obtained.

5.6.3 Ensure continuing processibility of Web sites and HTML pages

Web sites and HTML pages must be capable of replicating the content, context, and functionality without regard to technology changes for as long as necessary. This capacity for replication is known generally in archival practice as maintaining processibility. Processibility means, among other things, that Web sites and HTML pages can be moved from one technology platform to an entirely different one without any loss of data integrity and little or no degradation in the capability to render the original presentations. In some instances it may not be possible to fully reconstruct the physical layout of every Web site or HTML page, especially when proprietary plug-ins or special features are used. Nonetheless, even partially reconstructable Web sites and HTML pages are processible and should be capable of being read and interpreted correctly with current and future technologies for as long as necessary.

Web sites and HTML pages transferred to an electronic archives repository will have an indefinite retention period. During this long retention period storage media and the devices used to read them will fall victim to technology obsolescence. In order to ensure the continuing processibility of HTML pages and Web sites, the storage media on which they and associated metadata are stored must be renewed. The frequency with which this renewal may occur depends upon the error status of the current storage media and the rate of obsolescence of the existing media and drives.

The experience of the National Archives of the United States in transferring electronic records to new storage media suggests that this should be done every ten years (absent any evidence of a pending catastrophic media failure that a media monitoring program is likely to reveal). Transfer of Web sites and HTML pages along with associated metadata to new storage media every ten to fifteen years should be sufficient, barring any exceptions related to the chosen media or file format.

There are two established ways to renew storage media. One is to copy the bit string comprising Web sites, pages, and documents from an existing storage medium to a newer version of the same medium (3480 cartridge tape to 3480 cartridge tape). A copy therefore reproduces an exact mirror image of the bit stream on the new medium. The second way is to reformat the bit string comprising Web sites and HTML pages on one storage medium to one that is different (e.g., 3480 cartridge tape to Digital Linear Tape).

Recommended Best Practice:

- Extend the processibility of Web sites, HTML pages, and Web documents by copying and/or reformatting them to new storage media every 10 to 15 years.

5.6.4 Preserve content and contextual integrity of Web sites and HTML pages

Preserving the content integrity of Web sites and HTML pages means that no undocumented substantive alteration or loss in information content occurs during storage or during the migration of the data to new technologies for ensuring continuing processibility. Preserving the contextual integrity of Web sites and HTML pages involves maintaining metadata such as historical logs and an audit trail of all actions taken to extend their processibility, namely when the action was done, how it was done, why it was done, who performed it, and the results. This is known as preservation metadata. In some instances it may not be possible to preserve the exact appearance, content and functionality of Web sites and HTML pages, especially when browser dependent tags, proprietary plug-ins, and third-party Web sites are involved. Any loss of functionality, content, or appearance should be fully documented in preservation metadata.

Protecting the internal integrity of a Web site includes ensuring that internal links within the site remain intact when the technology platform changes, as in the case of migrating a Web site and associated HTML pages to an electronic archives repository. The easiest way to protect these links and avoid a time consuming reconstruction of a Web site directory is to employ relative URL addresses.

Web sites and HTML pages, like other electronic records, are vulnerable to some loss of data over time through media degradation and preservation activities to extend their usability through conversion to a new storage format. Protecting the integrity of Web sites and HTML pages in this context requires implementation of quality control procedures that generate bit/byte counts, Cyclical Redundancy Checksums, or hash digests that can be compared with bit/bye counts, Cyclical Redundancy Checksums, or hash digests generated at some point in the future.

Recommended Best Practices:

- Design and implement methods, such as checksums or hash digests of individual HTML pages or a public Web site as a whole, that provide a means for detecting any change in the informational content of the pages or site and preservation metadata,
- Employ relative URL addresses to ensure that internal Web links remain intact when technology platforms change,
- Document any loss in appearance, content, or functionality of Web sites resulting from the use of extended HTML tags or third-party Web sites that are no longer operational,
- Compute a cyclical redundancy checksum or hash digest of each Web site and HTML page after each preservation activity (e.g., reformatting to new media) and prior to the next preservation activity, then compare them to establish that there has been no alteration in the underlying bit stream that comprises the data, and

- Maintain an audit trail of the preservation history of each Web site including cyclical redundancy checksums and hash digests.

5.6.5 Employ XHTML as a vendor technology neutral format

Virtually all software vendors develop native, proprietary file formats to maximize performance and to enhance analytical techniques. These proprietary formats require the use of specific application software and sometimes a specific hardware/software technology platform. This dependence can be overcome through the use of open, vendor technology neutral data interchange formats. "Open" refers to the specifications for the format being in the public domain and therefore the data can be used with any vendor software product that imports data in this format. A vendor technology neutral format with a wide base of vendor implementations, therefore, would support Web sites, HTML pages, and associated metadata independence across multiple technology platforms and applications over time.

To put it differently, data independence means when there is a reference request for a particular Web site or HTML page and associated metadata at some point in the future they could be easily ported to any system platform or application that supports the format. Of course, over an extended period of time, say, thirty years or so, it is possible that a format may become obsolete and it may be necessary to migrate the data to the newer, more widely supported format.

Currently, the W3C is promoting "data independence" of Web sites and HTML pages through its support for XHTML as a vendor technology neutral "open" format. The proponents of XHTML believe that this strong base of support will encourage the developers of software tools for Web sites and HTML pages to ensure that their products conform to XHTML. The support of W3C for XHTML, the availability of multi-vendor products that conform to XHTML, and a wide base of user support suggest that XHTML and its variants are likely to persist over time.

It is likely that most of the Web sites and HTML pages transferred to the SIA electronic archives repository will be 4.0. At the time of transfer or shortly thereafter they should be migrated to XHTML in order to maximize "data independence" and to standardize, where possible, browser dependent tags and the inappropriate combination of HTML tags.

Recommended Best Practices:

- Adopt a policy that mandates the use of XHTML for all Web sites and HTML pages resident on a SI public Web server, and
- As appropriate, migrate all Web sites and HTML pages to XHTML when they are transferred to the electronic archives repository

5.6.6 Select a durable and cost-effective off-line storage medium

The selection of a durable and cost-effective storage medium for the archival preservation of Web sites,

HTML pages, and associated metadata is a threshold issue because it is likely that some form of off-line storage will be used. Two recording technologies (magnetic and optical) currently are used for off-line storage. There are three different magnetic recording technologies in use today - longitudinal, serpentine, and helical. Optical recording technologies include read only (ROM/DVD-ROM), write once, read many times (WORM/DVD-R), and rewriteable (RW). There is a great deal of debate about which recording technology is appropriate for use in an archives where electronic records and data must be retained indefinitely. Criteria for selecting a storage medium could include: high storage capacity; high data transfer rate; a predicted life expectancy of at least 20 years; established and stable market place presence; and suitability. (52)

Recommended Best Practices:

- Adopt the following criteria for use in selecting a storage medium to be used
- High storage capacity
- High data transfer rate
- Predicted life expectancy of at least twenty (20) years
- Established and stable marketplace presence
- Conduct a risk analysis of the tradeoffs between optical and magnetic storage media
- Select the storage medium that minimizes long-term risks

5.6.7 Maintain a secure electronic archives repository

Protecting Web sites, HTML pages, and associated metadata from alteration, deletion, or loss is a critical component of archival preservation that involves several different but related best practices. The first line of defense against alteration or loss of Web sites, HTML pages, and associated metadata is their transfer from an operational environment to a repository/archives where access to the data and the manipulation possibilities for the data can be strictly controlled. Such controls generally allow the data to be accessed, viewed, printed and copied, but not changed in any way. Protecting the data against alteration, deletion, or loss also should include locating the electronic archives repository in an area where the threat of a natural disaster such as a flood or earthquake is minimal. A disaster recovery plan should be in place that can be quickly implemented in the event of a natural calamity. In addition a "fail safe" recovery mirror copy of all Web sites, HTML pages, and associated metadata in the electronic archives repository should be maintained at a second geographical location.

The off-line storage media for electronic records are magnetic and optical storage, which are inherently fragile so their long-term readability and longevity are at risk. John Van Bogart of the National Media Laboratory (NML) has documented that temperatures in excess of 26º C (74º F) and a relative humidity of 70 percent can significantly reduce the predicted life expectancy of magnetic and optical storage media.

Recommended Best Practices:

- Maintain XHTML pages and metadata in secure environment where read-only access is controlled.
- Create a mirror copy of all Web sites and associated metadata and store it at a second electronic archives repository that is geographically removed from the primary archives/repository.
- Have in place a disaster recovery plan in the event of a natural disaster such as a flood, fire, or earthquake.
- Maintain a stable temperature of 15º C (59º F) and 40 percent relative humidity for the storage and management of the media and associated data.

5.6.8 Implement a quality assurance media monitoring program

A media-monitoring program should be in place that consists of a periodic (preferably annually) inspection of selected storage media to ensure that no catastrophic loss has occurred or is impending. (53) The findings and any remedial action taken to correct deficiencies should be documented in preservation history metadata.

As noted earlier, magnetic and optical media are the two digital storage media widely used today for archival storage. Although both classes of storage media are robust they are inherently fragile and therefore their durability and readability are at risk. Research conducted by the National Media Laboratory on predicted life expectancy of magnetic and optical storage media demonstrates that lower temperatures and lower relative humidity levels can contribute significantly to their life expectancy. (54) There is no guarantee that lower temperatures and lower relative humidity levels alone will ensure that storage media remain durable and readable over their entire predicted life expectancy.

Recommended Best Practices:

- Conduct an annual inspection of a sample of storage media to ensure that no catastrophic loss has occurred or is impending.
- Document all findings resulting from the annual inspection, including what corrective action, if any, was taken in the preservation metadata.
- Identify storage media with 10 or more temporary read "errors" and copy them to new storage media of the same type (e.g. 3490 tape to 3490 tape) or reformat them to new but different storage media (e.g., 3490 tape to DLT tape).

5.6.9 Migrate Web sites, HTML pages and associated metadata as often as necessary to avoid technology obsolescence

The Web sites, HTML pages, and associated metadata transferred to the electronic archives repository will be preserved in processible form for as far into the future as possible. It is inevitable that new technology platforms will become available that support newer and more efficient vendor technology neutral formats that offer significant data management and access improvement. Establishing a time frame within which to migrate Web sites and associated metadata to a new and more efficient vendor

technology neutral format is difficult partly because the date at which a format may become totally obsolete is unknown. The commitment of W3C to the support of XHTML and the growing vendor development of tools to support it augur well for its longevity and for backward compatibility with any new vendor technology neutral formats.

Recommended Best Practices:

- Adopt XHTML as the preferred vendor technology neutral format for archival storage of SI Web sites and HTML pages, and
- Periodically migrate SI Web sites HTML pages, and associated metadata to updated or new vendor technology neutral formats

5.6.10 Create and maintain an historical preservation log

The long-term preservation of accessible and trustworthy Web sites and HTML pages requires the collection and maintenance of metadata that both identifies their profile attributes to support timely retrieval and describes the storage, maintenance, and migration circumstances that can help guarantee their trustworthiness. The Preservation Metadata Model in Appendix 2 identifies the kinds of information that should be captured and maintained in an Historical Preservation Log.

Recommended Best Practices:

- Establish and maintain an Historical Preservation Log that tracks all of the preservation activities undertaken to extend the usability and trustworthiness of Web sites and HTML pages.
- Incorporate the data elements in Appendix 2 into the Historical Preservation Log.

SOURCES

Charles Dollar, Authentic Electronic Records: Strategies for Long-Term Access (Chicago: Cohasset Associates, 1999).

Stephen Harries, "Capturing and Managing Electronic Records from Websites and Intranets in the Government Environment," proceedings of the LM Forum '99 European Citizens and Electronic Information: The Memory of the Information Society (Luxembourg: Office for Official Publications of the European Communities, 2000): pp. 72 - 80.

Gregory S. Hunter, Preserving Digital Information: A How To Do It Manual (New York: Schuman Publishers, 2000)

IM Forum- Internet and Internet Working Group, An Approach to Managing Internet and Intranet Information for Long Term Access and Accountability. September 1999. Available electronically at **http://www.imforumgi.gc.ca**.

IM Forum-Internet and Intranet Working Group, Managing Internet and Intranet Information for Long Term Access and Accountability - Implementation Guide. September 1999. Available electronically at **http://www.imforumgi.gc.ca**.

William LeFurgy, "Records and Archival Management of World Wide Web Sites," The Quarterly, The DC Caucus of the Mid-Atlantic Regional Archives Conference Vol. 2 No. 4 (March 2001).

Tanya Marshall, "Web-Based Records and Electronic Records Management: A Case Study with the National Museum of American History," (Intern Study, Smithsonian Institution Archives, December 1999).

Charles McClure and Timothy Sprehe, "Guidelines for Electronic Records Management on State and Federal Agency Websites." February 1998. Available electronically at **http://www.istweb.syr.edu/~mcclure/guidelines.html**.

Susan McKinney, "Managing Web-Based Information In a Changing Environment," 1998 Managing Electronic Records Conference (Chicago, 1998).

Mississippi Department of Archives and History, "Electronic Records Draft Guidelines. Part 5: Webpages," (Mississippi Department of Archives and History). Available electronically at **http://www.mdah.state.ms.us/arlib/erglweb.html**.

National Archives of Australia, Archiving Web Resources: A policy for keeping records of web-based activity in the Commonwealth. Revised January 2001. Available electronically at **http://www.naa.gov.au/recordkeeping/er/web_records/intro.html**.

National Archives of Australia, Archiving Web Resources: Guidelines for keeping records of web-based activity in the Commonwealth. January 2001. Available electronically at **http://www.naa.gov.au/recordkeeping/er/web_records/intro.html**.

National Archives and Records Administration, "FAQ on Technical Issues-Revised Snapshot of Agency Public Web Sites." Available electronically at **http://www.nara.gov/records/faq-tech.html**.

Johanne M. Pelletier, "Recordkeeping the Web: Challenges for Records Managers and Archivists (Part I)" and Garron Wells, "Recordkeeping the Web: Challenges for Records Managers and Archivists (Part II), papers presented at the Annual Meeting of the Association or Records Managers and Administrators (Las Vegas, Nevada, October 2000).

Seamus Ross, Changing Trains at Wigan: Digital Preservation and the Future of Scholarship NPO Preservation Guidance Occasional Papers (London: National Preservation Office, 2000)

APPENDIX 1   DOCUMENTATION OF SI WEB SITES AND HTML PAGES

The Smithsonian Institution Archives (SIA) has adopted an overall plan and strategy for the preservation of usable and trustworthy Smithsonian Institution Web sites and HTML pages of archival value that involves transfer of Web sites and HTML pages to an archives Web server. A key component of this overall plan and strategy is the collection and maintenance of technical information about Web sites and HTML pages that will help identify the technology context of the sites and pages and instill user confidence in their trustworthiness, especially if some functionality of a Web site or HTML page can no longer be supported.

In the context of this data collection effort a Web site refers to a home page that is owned and managed by the Smithsonian Institution (e.g., SI Main/Home Page) or an operational organizational unit within the larger organization such as the National Museum of American History. A Web site is the first thing users see when they enter the site, and typically it contains a site map and links to various Web subsites or HTML pages that comprise the Web site.

1. What is the Web site name?
2. What is the name of the Web site manager?
3. When was the Web site created?
4. How many times has the Web site undergone major redesign?
5. When was the most recent major redesign?
6. How many Web sub-sites and HTML pages currently made up this Web site at the time of transfer to the archives Web server?
7. What is the name or URL of the oldest HTML page?
8. What is the name or URL of the most recent HTML page?
9. Have any HTML pages been retired from active use?
10. Where are these "retired" HTML pages stored?
11. What operating system and hardware supported the Web site?
12. What is the name of the database (if any) that that supports the Web site?
13. What Web plug-ins are used in the Web site?
14. What is the name of the search engine (including version number) used in the Web site?
15. Which HTML pages have dynamic features that require interaction with viewers?
16. Which HTML pages involve the use of Active Server Page technology?
17. Which HTML pages have third-party Web site links (as defined above)?
18. Is the Dublin Core or an adaptation of it used to assign profile attributes to each Sub site or page?

19. Which version of HTML was used in designing the Web site or pages?
20. As a general rule who designs and maintains the Web site and HTML pages? SI staff? Contractor?
21. Have any Web sites been migrated from an earlier version of HTML to a more recent one?
22. Is there a version control functionality (e.g. change log) in place that captured information about substantive revisions in a Web site or HTML pages?
23. Are copies of a Web site created and maintained?
24. What means or tools are used to protect Web site and HTML integrity when generating a back-up copy or migrating to a new format or storage medium?

APPENDIX 2   METADATA PRESERVATION MODEL

Preservation Metadata Model for Web Sites and HTML pages

The metadata requirements for tracking and preserving Web sites and HTML pages that are identified in this Appendix draw upon recordkeeping principles and requirements, best archival practices, and the Dublin Core. Specifically, these requirements incorporate the guidelines, recommendations, and best practices identified in the Public Record Office Victoria (Australia) VERS Metadata Scheme, (PROS 99/007 Specification 2), the University of British Columbia study "Protecting the Integrity of Electronic Records," the University of Pittsburgh "Metadata Specifications Derived from Functional Requirements: A Reference Model for Business Acceptable Communications, DOD 5015.2 "Design Specifications for Electronic Records Management Software Applications, and 36 CFR Part 123 (National Archives). These requirements are organized into three areas: (1) A General Description of the Format; (2) Web site and HTML page identification data; and (3) Preservation data for each Web site and HTML page.

1. Metadata Format Description

This metadata area consists of a description of the overall format of the Tracking and Preservation Model Web sites and HTML pages of archival value. Its purpose is either to provide information or point to other information sources that will enable users in the future to understand the format of the records and their content. The date format followed is that specified in ISO 8601.

One way to satisfy this requirement for a metadata format description would be a short paragraph that identifies the SI policy and procedure under which the metadata format is constructed. This paragraph also should include a statement of how the information in the metadata is represented. An example of such a description would state that the metadata was:

Produced according to the SI Web site and HTML page Long-Term Access Policy, version X.X, mm/dd/yy. The structure of this record is represented using Extensible Markup Language (XML), 1.0, W3C,

2000.

2. Identification Data for

- **Web site/page name**
- **Content owner**
  - Organizational unit primarily responsible for management of the Web site/page
- **Date of creation**
- **Date became inactive**
- **Databases used**
- **Plug-ins used**
- **Browser version required**
- **Other proprietary software used**
- **Third-party Web sites**
  - Original URL
  - Date URL link was broken
  - Internal SIA URL with documentation about the third-party Web site/page

3. Preservation Data

a. General

- **Date of transfer to the SI electronic archives repository**
- **Format of Web site/pages**
- **Certification of valid data transfer**
  - Physical record count
  - Record integrity
    - CRC
    - Hash digest
- **Software support for file format**
  - Vendor
  - Version

b. Identification and Location of Storage Media

- **Type of electronic medium (e.g., magnetic)**
  - Physical location
    - Primary site
    - Secondary site

c. Reformat to new storage media

- **Date of reformatting**
- **Reformatting iteration number (repeatable)**
  - Storage medium
  - Storage location
  - Primary site
  - Secondary site
  - Physical record count
    - Before reformatting
    - After reformatting
  - Authentication of record integrity
    - CRC
    - Hash digest
  - Certification of accurate reformatting
    - Discrepancies (if any)
    - Corrections (if any)

d. Copy to new storage media

- **Date of copying**
- **Copying iteration number (repeatable)**
  - Storage medium
  - Storage location
  - Primary site
  - Secondary site
- **Physical record count**
  - Before recopying
  - After copying
- **Authentication of record integrity**
  - CRC
  - Hash digest
- **Certification of accurate copying**
  - Discrepancies (if any)
  - Corrections (if any)

e. Migration to new technology neutral file format

- **Name of new technology neutral file format**
- **Software support for technology neutral file format**
  - Software vendor
  - Software version
- **Date of migration**
- **Identification of layout, content, or functionality lost in migration**
- **Migration iteration number (repeatable)**

- ❍ Storage medium
- ❍ Storage location
    - ■ Primary site
    - ■ Secondary site
- **Physical record count**
    - ❍ Before conversion
    - ❍ After conversion
- **Authentication of record integrity**
    - ❍ CRC
    - ❍ Hash digest
- **Certification of accurate migration**
    - ❍ Discrepancies (if any)
    - ❍ Corrections (if any)

Contact us at osiaref@osia.si.edu

SIA Home || **Institutional History Division** || **National Collections Program**

Smithsonian
Institution

Revised: October 18, 2001

Smithsonian Institution Archives

Notes

(1) This study incorporates the definition of trustworthy records in "Records Management Guidance for Agencies Implementing Electronic Signature Technology" issued by the National Archives and Records Administration in October 2000 (Available electronically at http://www.nara.gov/records/policy/gpea.html). This definition identifies three characters of trustworthy records. Trustworthy electronic records are reliable because they are a full and accurate representation of events and activities and can be depended upon as memorials of them. Trustworthy electronic records are authentic because they are what they purport to be, that is, the individual or organization claiming to have created or received a record in fact did so. Trustworthy electronic records possess integrity because they are complete and unaltered over time as they are reformatted or copied to new storage media, converted to new technology platforms, or migrated to vendor technology neutral formats.

(2) This estimate is based upon an analysis of the Smithsonian Home Page and to various Web subsites that can be accessed through hypertext links.

(3) Two exceptions are the live video feeds of the two Pandas at the Zoo and daily live animal demos and the "Live Web Cam" of the Garber Restoration Facility.

(4) I discussed the concept of rescuing electronic records in danger of being lost through neglect in a paper 'Trends in the Archival Acquisition and Preservation of Electronic Records, 1970 - 2000," presented at conference on Archives and Cyberspace: East Meets West held at Moscow State University, Moscow, Russia, January 4-6,1996.

(5) Seamus Ross notes that preservation features should be made an 'integral part of the initial design of systems and projects." Changing Trains at Wigan: Digital Preservation and the Future of Scholarship NPO Preservation Guidance Occasional Papers (London: National Preservation Office, 2000): 13.

(6) "Waiting for Ghost Trains: Strategies for Managing Electronic Personal Records Before It Is Too Late," Archival Issues, Vol. 24, No. 1. An electronic version of this paper may be found at http://www.rbarry.com.

(7) The Dublin Core Metadata Element Set consists of fifteen data tags or elements used to describe information resources that will support consistent retrieval of these resources in a networked environment. The Dublin Core constitutes a common set of elements whose semantics are widely understood and implement, international in scope, and extensible to other metadata initiatives such as the Resource Description Framework (RDF). The fifteen Dublin Core elements are: Title, Creator,

Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, and Rights. For more information see http://www.dublincore.org/documents/2001/04/12/usageguide/.

(8) There was a 3.0 version with a number of new features that were never adopted. Some of these new features subsequently were incorporated into 3.2 and 4.0.

(9) For more information see "Clean up your HTML pages with HTML Tidy." Available at http://www.w3.org/People/Raggett/tidy/. October 31, 2000.

(10) A list of these extensions is available at http://ei.cs.vt.edu/~wwwbtb/book/chap13/nn2add.html. November 10, 2000.

(11) Tom Savola, Special Edition Using HTML (Que Corporation: Indianapolis, IN, 1995), pp. 180-181.

(12) This point is made in Archiving Web Resources: Guidelines for keeping records of web-based activity in the Commonwealth Government (National Archives of Australia, March 2001), p.8. Available electronically at http://www.naa.gov.au

(13) Even though the compression algorithm in MP3 (MPEG-1, Layer III) substantially reduces the volume of data, it still represents about one (1) Megabyte per minute of sound recording.

(14) They identify eight data elements: Title or name of posting, Version number of posting, Originating author, Hyperlinks in the posting, Date of initial posting, Date of last modification, Date of replacement/withdrawal, and Disposition. Guidelines for Electronic Records Management on State and Federal Agency Websites, p. 12.

(15) A case in point is the following scenario. When a user clicks a "Submit" button on a form or a hypertext link on a page, the user's Web browser sends an HTTP request to the Web server via the Internet or an Intranet. The Web server passes the data submitted and the appropriate page to a server such as ColdFusion through a server API. The server interacts with database servers, file systems, E-mail, EDMS, and other applications and retrieves the specified data. Then the server dynamically generates an HTML page that is returned to the Web server. The Web server returns the HTML page to the user's browser for viewing.

(16) Windows NT supports the Tar utility in Arcserve, Backup Exec, and NT Backup. Novell supports the Tar utility in Arcserve, Backup Exec, and Sbackup. See "Technical FAQ for NWM 05-01, Snapshot of Agency Public Web Sites. It is available electronically at http://www.nara.gov/records/faq-tech.html.

(17) One alternative to executable snapshots is to capture an entire Web site as PDF images with active links. See Ted Alspach, PDF with Acrobat 4 (Berkley, CA.: Peachpit Press, 1999), pp. 143 - 150.

Acrobat, of course, is a proprietary software application and lacks an export functionality, which means that a Web site and associated HTML pages cannot be exported out of PDF back into HTML, although it can be exported in ASCII. For further discussion of PDF see Charles Dollar, Authentic Electronic Records: Strategies for Long-Term Access (Cohasset Associates: Chicago, 1999), pp. 184 - 186.

(18) See "MEMORANDUM TO CHIEF INFORMATION OFFICERS: Snapshot of agency public web sites," National Archives and Records Administration, January 12, 2001.

(19) Archiving Web Resources: Guidelines for keeping records of web-based activity in the Commonwealth Government, National Archives of Australia (March 2001). It should be noted that under these guidelines a snapshot is one of several alternatives for capturing Web sites.

(20) Seamus Ross and Ann Gow, Post-Hoc Rescue of Digital Material (London: British Library and Joint Information Committee, 1998).

(21) I discussed this concept in a paper "Archivists and Librarians: Our Common Ground in the Information Age," presented at the 4th Conference of Librarians, Archivists, and Documentalists, Braga, Portugal, March 4 - 6, 1992

(22) Jeff Rothenberg, Using Emulation to Preserve Digital Documents (The Hague: National Library of the Netherlands, 2000), p. 24.

(23) For a balanced and thoughtful assessment of the Rothenberg proposal see Stewart Granger, "Emulation As A Digital Preservation Strategy," D-Lib Magazine (October 2000), Vol. 6 Number 1.

(24) The United States National Science Foundation and the Joint Information Systems Committee are supporting a join project by the University of Michigan and Leeds University (UK) called CAMILEON, which stands for Creative Archiving at Michigan & Leeds. For more information on the project see http://www.si.umich.edu/CAMILEON/.

(25) Preserving Digital Information: Report on the Task Force on Archiving of Digital Information, Commissioned by the Commission Preservation and Access and the Research Libraries, Inc. (Washington: Commission on Preservation and Access, 1996). I follow generally the Migration strategy advocated in this study but disagree about some of the details. For more information see Charles Dollar, Authentic Electronic Records: Strategies for Long-Term Access (Cohasset Associates: Chicago, 1999).

(26) For a strong defense of the Migration Strategy see David Barman, "Reality and Chimeras in the Preservation of Electronic Records, D-Lib Magazine (April 1999), Vol. 5 Number 2.

(27) This number is approximate because some museums (e.g., the National Air and Space Museum) maintain a public web server that has a link to the SI Public Home Page. Mike Tuttle, the SI Webmaster, reported that the SI public Web server has about 13 Gigabytes of HTML pages. Interview, February 20,

2001.

(28) "A Report of the Smithsonian Webmaster: The State of the Smithsonian Homepage July 18, 2000," [Electronic File, October 16, 2000): 12.

(29) I am deeply indebted to Vicki Portway, NASM Webmaster, who shared her insights with me on this topic.

(30) These guidelines are available electronically at http://prism2.si.edu/webmasters/sercweb.htm [internal Smithsonian server; not available to the public].

(31) This document is available electronically at http://prism2.si.edu/webmasters/nasmops.htm [internal Smithsonian server; not available to the public].

(32) The full study is available electronically at http://smithson.si.edu:8080/dev/policy/sgl.html.

(33) Ibid.

(34) This design policy document is available electronically at http://prism2.si.edu/webmasters/sercweb.htm [internal Smithsonian server; not available to the public]. The term "relational" should be "relative."

(35) These guidelines are available electronically at http://prism2.si.edu/webmasters/silweb.htm [internal Smithsonian server; not available to the public].

(36) Vicki Portway, NASM Webmaster, reports that the entire Web site is currently being redesigned to make the NASM home page more user friendly and informative. Perhaps these features will receive greater visibility in the redesigned Web home page.

(37) Tanya Marshall, a doctoral student in the College of Library and Information Services at the University of Maryland, did a comprehensive study entitled, "Web-Based Records and Electronic Records Management: A Case Study with the National Museum of American History" (Smithsonian Institution, December 1999). Appendix C "Structure of NMAH Website" provides a very useful listing of all the HTML pages comprising the NASM Web Site.

(38) Filemaker Pro supports Open Data Base Connectivity (ODBC), an industry "standard" that facilitates the importing and exporting of databases that conform to ODBC (e.g., Microsoft ACCESS) and mitigates somewhat the exclusive reliance on Filemaker Pro.

(39) This definition incorporates language from National Archives of Canada, Information Management Standards and Practices Division, Record Keeping in the Electronic Work Environment --Vision (1996), p.1.

(40) T. R. Schellenberg, The Management of Archives (New York: Columbia University Press, 1965): pp. 67 - 69. In September 1999 the Canadian IM Forum published An Approach to Managing Internet and Intranet Information for Long Term Access and Accountability. Implementation Guide (available electronically at http://www.imforumgi.gc.ca) in which a record refers to a "government record" that consists of "recorded information, regardless of its physical form, that a government institution controls and that is collected, created, received or used to initiate, conduct or complete an institutional activity. (p. 8) This same study defined publications as "information products" that are created and edited for the purpose of distribution or sale and are "compiled from available information or data sources that are published for a defined audience and a stated purpose." (p. 8). On the other hand, the Australian National Archives Policy on Archiving Web Resources - Revised January 2001 considers all Web sites and HTML pages to be "publication" because of their "inherently public nature" and that publications are also records (p. 11).

(41) This is the premise underlying the "accountability exposure" categories in Charles McClure and Timothy Sprehe, Guidelines for Electronic Records Management on State and Federal Agency Websites.

(42) For example, when the text of this report, which was created in Office 97, is saved in HTML and then viewed with Internet Explorer 5 browser the text is unchanged. However, the look and feel are substantially different. Pagination and footnotes are lost and different size type forms may be substituted. When a Word 97 document is converted (as opposed to imported) to FrontPage, a popular Web authoring tool, significant glitches can occur.

(43) It can be argued that when a Web-based record is identical in every respect to an electronic records or a paper-based record it may not be necessary to retain the Web-based record as long as the electronic or paper based record is captured in a recordkeeping system. See Archiving Web Resources: Guidelines for keeping records of web-based activity in the Commonwealth Government, National Archives of Australia (March 2001), p. 13.

(44) "Smithsonian Institution Archives Appraisal Methodology," (Smithsonian Institution Archives, September 1999): 11.

(45) A list of questions to be asked about "Electronic Records Created or Maintained" is prefaced with "Try not to think of electronic records as a record type, but as a record form." "Smithsonian Institution Archives Appraisal Methodology," 45.

(46) A list of production records of enduring value includes "Selected websites chosen according to importance and uniqueness." "Smithsonian Institution Archives Appraisal Methodology," 32.

(47) "Smithsonian Institution Archives Appraisal Methodology," 30 - 31.

(48) This does not mean that hypertext link between the SI Home Page and a museum center, for

example, should be broken.

(49) As noted elsewhere in this study, the use of Web plug-ins and proprietary search engines creates a software dependency that technology obsolescence exacerbates. In the long run, say fifty to one hundred years, the combined effects of software dependency and technology obsolescence may make it impossible to fully replicate the "look and feel" of Web sites and HTML pages created in the 1990s.

(50) Archiving Web Resources: Guidelines for keeping records of web-based activity in the Commonwealth Government, National Archives of Australia (March 2001), pp. 8 and 21 - 23. Note that this study uses both content driven and object driven to describe the same approach.

(51) "Smithsonian Institution Technical Reference Model Strategy, Smithsonian IT Management Briefing, April 2001.

(52) For an assessment of storage media see Charles Dollar, "Selecting Storage Media for Long-Term Access to Digital Records," The Information Management Journal (July1999), pp. 36 -43.

(53) For a discussion of a statistical sampling procedure that can be used to select the storage media see Dollar, Authentic Electronic Records, pp. 99 -100.

(54) According to John Van Bogart, the Principal Investigator for Media Stability Studies at the National Media Laboratory, 10º C (50º F) and 20% relative humidity provide the ideal storage environment for electronic records to be retained for long periods of time.

Contact us at osiaref@osia.si.edu

SIA Home || **Institutional History Division** || **National Collections Program**

Smithsonian
Institution

Revised: October 18, 2001