Smithsonian Institution Archives

SIA

June 2006

# CONSERVATION OF DIGITAL RECORDS

## *A Collaborative Electronic Records Project of the Rockefeller Archive Center and the Smithsonian Institution Archives*

Riccardo Ferrante
Information Technology Archivist

Presented at the 2006 Annual Meeting of the American Institute for Conservation of Historic and Artistic Works

# CONSERVATION OF DIGITAL RECORDS

# A Collaborative Electronic Records Project of the Rockefeller Archive Center and the Smithsonian Institution Archives

## Introduction

A great deal of study and effort has gone into the challenges of managing and preserving electronic records. The important role of digital records in the historical record is widely accepted. Much research in this area focuses largely on addressing state of the art technology, in other words, a focus on the future. However, some archives have been receiving digital records that date back several decades. Those records need not be lost in many cases. Proper handling of these records requires a three-fold response: preservation, conservation, and appropriate management. At the heart of this is the functional digital object, its authenticity, and its integrity.

Using the illustration of website conservation, this paper looks at the urgent need to address legacy formats, the risks digital records face, and the potential for effective solutions today.

Typically, records transferred to the Smithsonian Institution Archives have been inactive for at least three years. Some of the digital records in the Smithsonian Institution Archives collections are over thirty years old, reasonably referred to as ***legacy*** data formats. For example, personal papers acquired in 2006 are likely to include correspondence, manuscripts, or other artifacts on media ranging from 5.25 inch magnetic disks to computer hard disks.

The following is a brief case study of the Smithsonian Institution Archives' work in preserving and conserving websites to date and a description of its joint work with the Rockefeller Archive Center to implement digital preservation techniques.

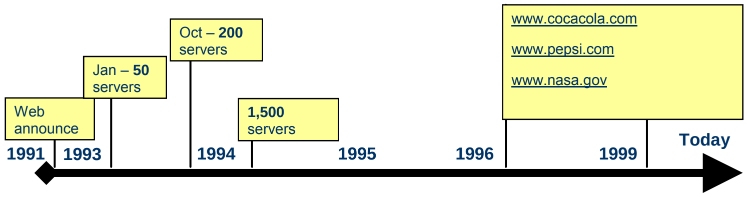
# Contents

# Timeline

## Adoption of the World Wide Web

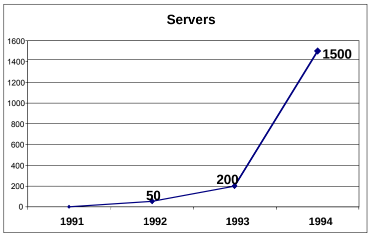
The World Wide Web, or Web as it is commonly referred to today, has revolutionized communication by governments, corporations, and the general public around the world. Today, global companies are competing to deliver new value as seen in Google's work to digitize and make available the contents of university libraries including Harvard, Stanford, Oxford, and the New York Public Library.

Local and regional governments are providing access to public records as well as conducting business through websites. Cultural heritage institutions are leveraging its technology to reach new audiences. Universities and colleges providing distance learning.  The impact of the Web on communication has been truly remarkable.



### *Growth Rate*

The extremely rapid adoption of this technology underscores the need to develop adequate conservation tools for digital object conservators. In just over a year since the world was introduced to the World Wide Web, 50 webservers were operational. Only ten months later, the number of servers grew to 200; a year later, webservers numbered 1,500 – 7 times the growth in the previous two years.



By June 2006

85.5 million

According to a survey of web servers completed in June 2006, web servers now exceed 85.5 million.[1] With new websites being established at such a rapid pace, conservation of digital objects is absolutely critical in order to retain this component of the historical record. Not only is conservation essential, but conservation techniques need to be sufficient to address the volume of valuable records being produced.

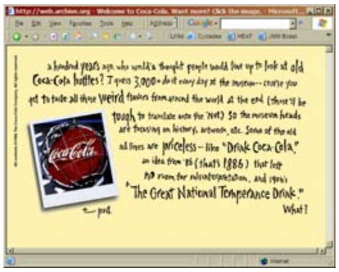
### *Exercising Care to Acquire Complete Records*

Some individuals quickly recognized the danger to the historical record posed by the speed with which this revolutionary tool was being adopted. One such person, Brewster Kahle, founded the ***Internet Archive*** to capture and preserve websites whose valuable information was made available over the Internet. This web-driven tool includes a search and retrieval component referred to as the ***Wayback Machine that*** allows a visitor to search for all websites offered under a particular website domain as defined by a specific URL[2].

Itself a software system that continues to mature and be enhanced, the completeness and accessibility of the websites archived at the Internet

[1] This figure represents distinct web domain names as reported in the June 2006 Netcraft Ltd. Web Server Survey.
[2] Universal Resource Locator

October 1996

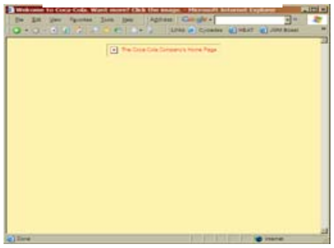


http://web.archive.org - Welcome to Coca-Cola. Want more? Click the image. - Microsoft...

a hundred years ago who would've thought people would line up to look at old Coca-Cola bottles? I guess 3,000+ do it every day at the museum-- course you get to taste all those weird flavors from around the world at the end (those'll be tough to translate onto the 'Net) so the museum heads are focusing on history, artwork, etc. Some of the old ad lines are priceless – like "Drink Coca-Cola," an idea from '86 (that's 1886) that left no room for misinterpretation, and 1906's "The Great National Temperance Drink." What!

← prd

June 1997



Welcome to Coca-Cola. Want more? Click the image. - Microsoft Internet Explorer

The Coca Cola Company's Home Page

Archive continues to improve as Internet technology advances.

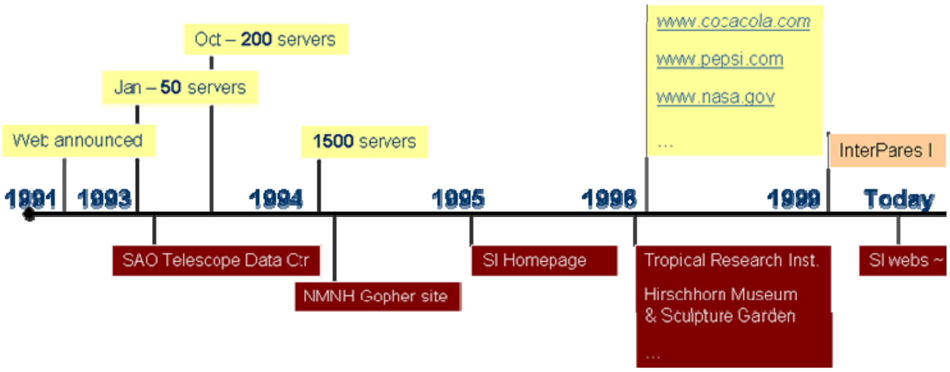
One factor affecting the completeness and therefore its integrity is the method used to transfer the digital object[3] into an archive's custody. If web-crawling technology is used to transfer a website, unlinked pages within the website are unlikely to be captured[4]. Sometimes referred to as the "dark Web" or the "semantic Web," they are not captured by web crawlers because the pages do not link to any page connected to the homepage of the website.

To the left is the October 1996 Coca-Cola website, the earliest version of http://www.cocacola.com captured by the Internet Archive. The third instance of this website as captured less than a year later demonstrates the impact of incomplete or flawed transfer methods. Similar examples are not limited to the early years of the web and are found in several archives.

Clearly, care needs to be exercised to acquire as a complete a record as possible. One must acknowledge that circumstances will occasionally prevent acquiring a complete digital object. However, the frequency of this problem can be significantly reduced if such care is exercised.

## Smithsonian Web Events

A term of James Smithson's bequest to the United States was the establishment of a museum devoted to the increase and diffusion of knowledge. In the larger context of the growth of the Web, the Smithsonian Institution had a number of noteworthy developments in its communication of that knowledge to the much wider constituency, now through the Internet.



The ***Smithsonian Astrophysical Observatory's Telescope Data Center*** website was launched in 1993, one of the first 250 websites on the Internet. It incorporated interfaces that allowed users to select different factors of observed data to generate graphs and other displays. It continues to be an active website more than a decade later, although a great deal of the original website and its code has been enhanced.

[3] In most cases, a complete website is considered to be a single digital object. Under some circumstances, a website may both be a single object and contain other digital objects. An example of the latter would be a website which included musical works.
[4] Some organizations may post one or more pages to the Internet through their website but isolate them from the website's main navigation paths for a variety of reasons.

Less than a year later, the National Museum of Nation History used the Internet and a different syntax, gopher, to post material for several scientific communities.

In 1995, the official website of the Smithsonian Institution was launched and screen shots of the website were included in the annual report to Congress. Over the next few years, more and more Smithsonian organizations established their own presence on the Web. In the same vein, commercial and other groups were similarly embracing the Web as a communication channel and awareness-building vehicle. Today, SI has more than 100 distinct websites available to the public, allowing them to visit the Smithsonian's museums and galleries "virtually."

## Three Reasons for Loss of Records

***Poorly managed objects***. A variety of factors contributes to the corruption or loss of records even when there is an appreciation of their part in the historical record. Where custodians or administrators of objects change frequently, it is common to assume that the digital objects in custody are complete and have been documented sufficiently to note if components are missing or non-functional.

This attitude is most often true of active objects. Since few administrators verify this aspect of the objects they inherit, a strong opportunity exists for components to be lost in the transfer of management. Several lesser factors contribute to this reason for record corruption and/or loss: changes in management procedures or object storage environments are two examples.

***Analog renditions lack authenticity***. Early solutions proposed for digital records focused on rendering the digital records to a media with which the archival community was already very familiar. For example, printing documents to paper or microfilming are still suggested for "born-digital" material.

This poses difficulties in that more and more born-digital records consist either in part or wholly of elements that cannot be rendered authentically through this approach. Audio, video, Macromedia Flash, or Active Server Pages components are a few examples. Where these elements can be included without loss of functional behavior, efforts must be made to conserve these items in a fully authentic manner. Further example of the difficulties are demonstrated by database-driven websites.

In light of well-respected leaders in the archival community embracing digital preservation and conservation such as the Library of Congress, the National Archives and Records Administration, the National Library of Australia, the British Library, CHIN, Cornell University, and others, the wider archival community has been slow to adopt digital preservation and conservation.

***Sophisticated dynamic elements***. The popularity of the Web has been driven by the IT community's increasing ability to develop sophisticated functionality accessible over the Web. Google's and Yahoo's search and retrieval sites are popular examples. But databases are not the only such components. Some sites offer magnification functionality on demand for visually impaired visitors that they can control. Other components may track the pages a person views and suggest other possible helpful pages or websites. Internet games are another. Dynamic elements like these are becoming more and more common in today's websites.

# The Collaborative Electronic Records Project

## Purpose and Guidelines

The Smithsonian Institution Archives and the Rockefeller Archive Center are completing the first phase of a three-year collaborative project to develop, test, and implement a technical system to preserve digital documents. As part of the project, we are sharing our findings with other nonprofit organizations through a number of different channels. The project outputs will include a generic system model, email guidelines, transfer guidelines, and a sample business case and implementation project plan.

Nonprofit archives face a special challenge to address digital records in light of their limited resources. A large portion of this record, deemed vital to understanding the history of individuals and organizations, is being lost before it even crosses the threshold, lost while the archival profession has struggled to find viable strategies for preserving the records' authenticity, integrity, reliability, and accessibility.

As fellow nonprofit archives, the Smithsonian Institution Archives and Rockefeller Archive Center find it essential to make practical headway to retain and preserve this growing segment of the historical record. Given modern digital forms of information, the long-term preservation of electronic records, particularly email, will be critically important for scholars looking at the first decade of the 21st century, as well as for organizational accountability. For this reason, the Collaborative Electronic Records Project has chosen email records as its focus.

## Participating Archives

### *Rockefeller Archive Center*

The Rockefeller Archive Center serves as the permanent repository of records of several philanthropic organizations in addition to Rockefeller-sponsored foundations, trusts, and family records of the Rockefellers for a total of more than twenty distinct depositors. The Rockefeller Archive Center does not have any direct record management responsibilities with its depositors. However, it is occasionally consulted on record management issues.

### *Smithsonian Institution Archives*

The Smithsonian Institution Archives is the repository for historical records of the Institution itself. The mandate for the Archives includes the role of Record Manager for the Institution. In this capacity, the Archives works with the different units to establish and update records disposition schedules. The Archives accessions and manages the official record series designated with limited retention.

## The Value of Collaboration

The combination of these two archives brings together their unique experiences described above and thereby enables them to represent a broad spectrum of organizations. It is the goal of the joint team that the applied results of the project will assist peer organizations in the area of digital record conservation and management.

# Examples of Website Conservation Issues

## Coding variations

Navigation through a website can be encoded in numerous different manners. This tends to occur most frequently when a website has had multiple webmasters either in succession or concurrently.

From its earliest versions, HTML did not require a specific syntax for its URL encoding. To enter a website, one has always needed a full URL address, e.g. http://www.si.edu/archives/, whether entered by hand into the web browser or embedded on a webpage elsewhere on the Internet.

However, hyperlinks to pages within a website could by encoded using absolute path addresses, e.g.,

```
<A HREF>http://www.si.edu/index.htm<a>
<A HREF>http://www.si.edu/exhibitsandprograms.htm<a>
```

Equally viable for URL encoding of links between these two pages is ***relative paths***, which reflects where pages are found on the webserver in relationship to each other. For example on the index.htm page, a hyperlink to Exhibits and Programs encoded as <A HREF>exhibitsandprograms.htm<A> assumes that both the index.htm and the exhibitsandprogram.htm files are located in the same directory on the webserver. A visitors page located in the subdirectory “info” would be encoded on these pages as

```
<A HREF>info/visitors.htm<A>
```

To navigate from the visitors page back to the index page, a relative path encoding would be structured as:

```
<A HREF>../index.htm<A>
```

The **../** instructs the web browser to look in visitors.htm’s parent directory for the index.htm file.

```
🗀 index.htm
└ 🗀 info/visitors.htm
```

A third encoding possibility for early websites is a combination of the two methods already discussed, referred to as “absolute relative” paths. Absolute relative paths specify the full path and filename of the webpage in relationship to the root directory of the website. The webserver name, e.g. www.si.edu, is not included in this type of URL encoding.

A website that consistently uses relative path hyperlink encoding throughout will yield a largely, if not fully, functional collection object without any immediate conservation concerns in the area of hyperlinks. A researcher can be given access to a reference duplicate of the object on a computer without any access to the Internet and experience without any negative impact on their ability to study the website.

A website that employs absolute path or absolute relative path methodology for its hyperlink encoding will fail to deliver the original functionality of the full website object in original bit stream order. Conservation is required if the functionality of the original object is to be maintained. Conservation ***is possible*** for the issue described here without violating the best practice commitment to open standards-based protocols. In other words, conserving websites with absolute and absolute relative paths issues can be done while fully adhering to the current version of the open standard HTML.

### Dynamic Internal Components

Advancements in the technologies used as part of websites and webservers have resulted in dynamic components that “build” a webpage based on a set of embedded parameters. Other elements, such as streaming multimedia files, might be integrated into the website while physically residing on a different computer. Early elements include animated images, forms which generated output such as email, and scrolling content. The variety of dynamic components continues to grow at a rapid pace, requiring an ongoing need for conservation in this area.

### Dependencies on other systems

As technology advances, a digital object may be the construct of several interrelated systems. Database-driven websites are a simple example. One website may interface with several distinct databases on separate servers. These dependencies are best considered before this type of record is accessioned because each component system or data format expands the complexity of the preservation task, bringing its own unique conservation issues to the table.

## Guiding Values and Experiences

### Authenticity

A key value of the Rockefeller Archive Center and the Smithsonian Institution Archives is digital object authenticity preserving both the content and the behavior of the conserved object over time. These elements are considered essential for the expected scholars and other researchers.

### Viable Accessibility: Researcher Skills & Resources

Another key value is that these objects be accessible given a reasonable level of skill handling digital objects and given a reasonable level of available hardware and software. The archives consider it unreasonable to expect researchers to have the skills of an advanced webmaster as well as the hardware and software resources required, particularly over time. Therefore due consideration of appropriate conservation measures are necessary as the preservation tasks are determined.

## Best Practices

The application of best practices is essential. Best conservation practices exist to reduce the risk to the authenticity and integrity of records in order to extend the life of the record. Accessibility to the record is secondary to the paramount value of retaining the authentic record. Still, retaining the record directly serves to extend the record's life through accessibility to scholars and researchers.

The development of electronic records management best practices standards has been a major focus of the corporate and archival communities in the past several years. Both communities have the majority of these practices in common.

### *Choices of conservation formats*

Choice of conservation techniques is guided by the values described above regardless of record format. The techniques applied must be suited to the nature of the object.

Conservation of digital objects therefore must address the three obsolescence factors: file format, physical storage format, and the required environment for authentic functionality. One of the highest risks to data formats is the proprietary formats in which they are created.

These data formats are designed to reinforce users' dependency on the original application by the simple economics of the marketplace. Software vendors regularly create new enhanced versions to force the obsolescence of earlier editions of their applications for the purpose of growing acquired revenue.

Well-documented, open standards are essential to eliminate or reduce the dependency on proprietary applications. Standards developed by groups of users are designed to provide interoperability between sophisticated computer systems, software applications, operating systems, or hardware. Because standards are by their very nature shared definitions of structure and/or behavior, they typically work to extend the accessibility of a given digital object as compared to proprietary formats.

When data formats or technical environments are highly specialized, as is typically the case with proprietary formats, currently available conservation paths may be limited or non-existent. Still objects must be retained for their value in the hopes that standards and future technology will provide the tools to enable the continuing life of the authentic record.

### *Timing*

When official documents were exclusively in paper format, the central file was a standard feature of corporation's recordkeeping practices. As the personal computer spread throughout the business and government environments, the disciplined approach to recordkeeping was replaced with a less formal approach.

Official records, if digital, may be kept on individuals PC's or on the network. In either case, these files are often overwritten, being used as templates for updates or similar documents. Therefore, the early capture of the objects' original code becomes particularly useful in order to reduce the chance that the original record is lost or modified before transfer to the archive. This is a significant departure from standard accession methodology.

While some organizations have begun to implement PC's that prevent file storage on the PC itself, this does not remove the overwriting and deletion risk for the files stored on the network.

### Completeness

Because computer science technology has consistently demonstrated an increasingly rapid pace of advances in hardware and software, it is reasonable to project the possibility that in the future another conservation edition could be generated from the original code as another first generation edition rather than a second generation from the first conserved result. This course is clearly preferable and is the rationale behind our practice of storing an archival object consisting of the original code, the conserved code, and the associated metadata. In this manner we assure a complete yet conserved and accessible record.

## Findings to Date

### Starting Now Yields Valuable Results

Based on our experiences, important records already succumbing to one or more of the three obsolescence factors have already been conserved at the Smithsonian Institution Archives, yielding measurable successes. The conservation measures that can be applied to these objects today enable the archive to maintain an authentic scope of functionality otherwise impossible.

Since all the transferred digital records of an accession are assessed for preservation needs, the objects that do not have a suitable preservation format can be logged, assigning risk ratings, and preservation priorities. This valuable information helps to shape the conservation research agenda.

### Analog principles and paradigms

As the professional community has grappled with the challenges of digital records, we find that more and more of the principles and paradigms associated with analog material apply equally to the digital medium. This characteristic is a challenging aspect of the educational component as? part of establishing a successful digital preservation program.

### Curiosity and Persistence

Curiosity and persistence serve well as the conservation profession moves further into the area of digital works. Conservation tools already exist in a good many cases. At this time, some tools are "patchwork" affairs of combined applications.

The reality is that tools conceived and developed for other purposes make a surprising degree of digital conservation possible today. A curiosity that includes fields outside the direct field of conservation can result in major steps. For example, Brent Seales's (University of Kentucky) work to apply the principles of

CT scanning to objects whose conservation now requires no further human contact enables access to print objects.

### Future tools and tool enhancement

Seales's work also underscores the promise of technological innovations which will further the field of digital preservation and conservation. The increasing commitment to the use of open standards by software and computer vendors, and public demand for their incorporation in proprietary software systems and applications will help to develop a computing culture that is more sensitive to preservation concerns. Even digital objects which can only be retained as bit streams, or those where migration provides 'best editions' may well be conservable within a decade. But only if we start now.

For a partial list of conservation formats used by the Smithsonian Institution Archives and Rockefeller Archive Center, see Appendix 1.

# Bibliography

W. Brent Seales, James Griffioen, Kevin Kiernan, C. J. Yuan, and Linda Cantara. "The Digital Atheneum: New Technologies for Restoring and Preserving Old Documents."" *Computers in Libraries* 20:2 (February 2000), 26-30.

Netcraft Ltd. "June 2006 Web Server Survey." http://www.netcraft.com. (June, 2006)

# Appendix 1

| Object Type | Original Data Format | Conservation Format[†] | Object Type | Original Format | Conservation Format |
|---|---|---|---|---|---|
| | ASCII (Text) | No Change | | BMP | Same |
| ***Word Processing*** | | PDF/A | ***Presentation*** | | PDF/A |
| ***Web Formats*** | HTML<br>ASP<br>CFM | XHTML, XML | ***Audio*** | MP3, CDA | WAV |
| ***Images*** | GIF, PSD, PSP, PICT | TIFF | ***Audio*** | RM, RAM | None, unless unique functionality is not implemented. |
| ***Databases*** | Varied | XML, SQL | | | |
| ***Email*** | Varied | Body – ASCII<br>Attachments – see other categories | ***Video*** | ***Various*** | AVI |
| ***Computer-Aided Design*** | DWG, CAD | Vector-based PDF/A | ***Diagramming*** | VSD (Microsoft Visio) | PDF/A |

[†] The conservation formats identified here are those selected by the participating archives at the time of publication. Conservation format determinations are made a periodic basis, informed by technological advancements, standards development, professional best practices, and other information science developments.