DIGITAL PRESERVATION AT THE

U.S. GOVERNMENT PRINTING OFFICE:

WHITE PAPER

Version 2.0

9 July 2008

Record of Changes

Version Number	Description Of Change	Revision Date	Author
1.0	Baseline Document	July 12, 2006	George Barnum
2.0	Draft revision includes background on GPO's efforts to date and updated information on FDsys	May 23, 2008	Gil Baldwin
2.0	Additional revisions to incorporate reviewer comments on the version 2.0 draft	June 27, 2008	Gil Baldwin
2.0	Link added for Information Technology Glossary	July 3, 2008	Gil Baldwin

Abstract

This paper provides a non-technical introduction to digital preservation, particularly to the challenges for keeping digital content available and accessible into the future. Then the focus shifts to the digital preservation activities of the U.S. Government Printing Office (GPO) and highlights of GPO's technological transformation, moving from Web access, to content authentication, and development of a trusted repository. The final section introduces GPO's Federal Digital System, a content management system and digital repository designed to support GPO's mission of keeping America informed.

What is Digital Preservation?

Although media and methods for using and sharing information have changed rapidly, at least one fundamental requirement of preservation remains unchanged by technology: access and use are predicated on reliable availability of information. Information, regardless of its format or medium upon which it is recorded, must be preserved over time. Digital preservation is the sum of managed activities necessary for the long term maintenance of a byte stream and associated metadata sufficient to render a facsimile of the original document with its content and presentation intact, and to assure the continued accessibility of the content over time despite changing technology or data degradation.

One of the challenges of digital preservation has been to develop a common understanding of the term. Several definitions, in increasingly granular detail, have been developed by a working group within the Preservation and Reformatting Section of the Association for Library Collections and Technical Services (ALCTS), American Library Association. These definitions were published in 2008, and are intended to promote an understanding of digital preservation within the library community, its allied professions, and their user communities. The U.S. Government Printing Office (GPO) view of digital preservation is summarized in this ALCTS definition:

> Digital preservation combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time.

The Challenge

Documents printed on paper or film, whether primarily made up of text or images, have in the past composed the majority of information objects requiring preservation. Libraries and archives have developed approaches to preservation that assess an asset's condition, context (such as its place as part of a collection, its scarcity, or its intended impact or use), and priority and risk (what factors make this a more or less urgent candidate for preservation). Much of the emphasis in preservation of printed media is focused on assuring that the medium or substrate is stable and therefore persistent. This has included treating the native substrate or transferring the content to newer, more stable substrate.

Digital materials cause similar concerns about storage of data on media which may deteriorate or become obsolete. These concerns are typically addressed by copying, or refreshing, the data to a new storage medium. Digital content is also subject to concerns about the continued usability of the data over time as a consequence of rapid technological obsolescence. Digital

data depends on intermediary technology to be understood by humans. A single page of text displayed in a web browser may rely on many layers of computer code which instruct the computer to render the content consistent with the author's intended output. In the dynamic environment that has developed and advanced the technology of electronic communications, it is common for data formats, software, and storage media to be frequently replaced by newer applications and products.

For example, a document is created for use by the public over the Internet in a widely used, but proprietary, format. Over time new versions of the format are issued. Then the company that owns the rights to the format is sold. The new owner discontinues support for earlier versions of the format, and viewing the document becomes impossible. Although the data is intact and robust on its storage medium, it is no longer accessible.

Digital Preservation System Basics

Digital preservation involves multiple layers of data and processes, which must be managed in a coordinated fashion. At its most fundamental level, digital information is a stream of ones and zeroes. What collects and compiles it into meaningful information is intermediary technology: operating systems, applications, and data formats. In order to render the data meaningfully, intermediaries need additional data, or metadata, about the fundamental content.

Trust and the creation of trusted systems are the bedrock of digital preservation. In the absence of hard, visual markers which assure that the integrity of information is unimpaired and uncompromised, system designers are facing the challenge of providing structures on which such trust can be established and built.

The first and most basic structure for this trust is a clearly articulated and comprehensive reference model for managing preservation activities. A reference model establishes common terms and concepts, provides a framework for defining and describing functional entities, describes relationships between entities, and provides a foundation for analysis, evaluation, and standards development. GPO is employing the Open Archival Information System Reference Model (ISO 14721:2003), which defines the scope of designated communities of use, outlines the relationship of information producers and archivists, and effects sufficient control to ensure long-term preservation and access.

A successful digital preservation system will manage processes which:

- safeguard digital content data along with all relevant metadata,
- assess the condition and needs of collections of digital information,
- provide the method by which content is meaningfully rendered despite continually changing technology, and
- are auditable, replicable, and build the basis for trust.

Within the model, information is processed, stored, treated, and made available for access. Secure storage and reliable retrieval, access, and transmission are all areas in which the current information technology marketplace provides sophisticated, mature solutions. Systems

for performing preservation treatments or processes on digital files are still in their early stages of experimentation and development. Most authorities describe three fundamental activities for preserving digital information:

- Refreshment (or bit-level preservation) which assures that byte streams are transferred intact to newer or fresher storage media.

- Migration in which digital information created to function with a particular intermediary technology (such as a software application) is transformed or "migrated" to be able to function in newer or different version of that intermediary. Underlying information is retained but older formats and internal structures are replaced by newer.

- Emulation, in which a specially conceived system interprets what digital information requires to be rendered meaningfully and then creates or "emulates" that environment, even though the intermediary technology that previously did the work is now obsolete or extinct. The digital object is rendered in an environment that acts like the original.

A body of literature exists which defines these activities, but large-scale projects employing migration and emulation are scarce, and no clear consensus exists on preference of one over the others. It is assumed that any successful large-scale preservation program dealing with a wide variety of digital object types will employ multiple approaches, or combinations, depending on requirements and risks.

Highlights of Digital Preservation at GPO

GPO has statutory responsibility for producing Government information products and for disseminating Government information to the public on behalf of all three branches of the Government. One mission of the GPO is to provide, in partnership with the Federal depository libraries, for perpetual, free, and ready public access to the print and digital publications of the Government. GPO is developing its Federal Digital System (FDsys), which will allow federal content creators to easily create and submit content which will then be preserved, authenticated, managed and delivered upon request.

GPO documented its responsibility for permanent public access to Government information in digital form in the 1996 *Study to Identify Measures Necessary for a Successful Transition to a More Electronic Federal Depository Library Program*, a report to the Congress required by the Legislative Branch Appropriations Act, Public Law 104-53. The study recognized the implications of the digital age and the changing roles of GPO, the Federal agencies, Government information users, and the depository libraries. Within two years GPO had begun to acquire a collection of digital content, and in 1998 published *Managing the FDLP Electronic Collection: A Policy and Planning Document*, its initial plan to manage these digital resources.

In 2003 the National Archives and Records Administration (NARA) designated GPO as an official archival affiliate for the electronic content on GPO Access. The NARA-GPO memorandum of understanding provides for the permanent preservation and access to the online versions of the Congressional Record, the Federal Register, the Code of Federal Regulations, or other digital publications from GPO Access.

In addition to storing digital files on its own servers, GPO has entered into numerous partnerships with agencies and educational institutions to provide permanent access to digital content in scope for the Federal Depository Library Program (FDLP). An early example of this type of partnership is the Cybercemetery, a partnership between GPO and the University of North Texas Libraries to provide permanent public access to the Web sites and publications of defunct U.S. Government agencies and commissions.

In 2004, GPO reaffirmed its commitment to provide "perpetual, free, and ready public access to the printed and electronic documents …of the Federal government." This mission statement, articulated in *A Strategic Vision for the 21st Century*, established the foundation for the development of FDsys, GPO's system for digital preservation and access.

In early 2008, GPO launched its first authenticated databases live on *GPO Access*. For the first time, GPO digitally signed and certified the PDF files in the online federal budget released in February. The beta Authenticated Public and Private Laws for the 110th Congress database was incorporated into the live Public and Private Laws application on *GPO Access* in March. Both applications provide users with no-fee access to digitally signed PDF content. The digital signature provides assurance that an electronic document has not been altered since GPO disseminated it, verifying document integrity and authenticity of GPO online Federal documents.

Developing GPO's Trusted System

For many years GPO was predominantly a print-based environment in which the content on GPO Access was a by-product of the printing process. GPO is moving to a content-based environment, in which digital content may be created and submitted, and then be preserved, authenticated, managed and delivered upon request. Digital content will support printing in large or small production runs, as well as other outputs which users may request. One of the desired outcomes of this technology transformation is that GPO is widely recognized as a trusted provider of authentic, official Government information.

The first steps toward this goal focused on the delivery of trusted content, beginning with digital content designated as official and moving to authentication of digital documents. The next step in GPO's technology transformation is the development of a trusted system. In the introduction to their 2002 report on the attributes of trustworthy systems for preserving digital information, the Research Libraries Group (RLG) points out that

> All research resources need care and attention to survive, but digital research resources need more attention, often much sooner than resources on paper. The inherent fragility of digital materials leaves only a small window of opportunity to address this problem before ... [losing] resources on an ever-larger scale.

GPO is developing a digital content system capable of managing all known Federal Government documents within the scope of the FDLP and GPO's other information dissemination programs. FDsys is an integrated content management system which incorporates state-of-the-art technology for document authentication and digital preservation. Digital content in FDsys will be permanently available on the Web for searching, viewing, and downloading; for conventional and on-demand printing; or for other dissemination methods.

FDsys capabilities will be deployed in a series of releases. An internal proof-of-concept release of FDsys was completed in September 2007 to support beta testing. FDsys will become available to users in late 2008, beginning a process of incremental releases. Each release will add functionality to the previous one. The first public release will provide FDsys core capabilities, including such foundational elements as system infrastructure and security and a digital repository that conforms to the OAIS reference model and enables the management of content and metadata.

Digital preservation in GPO's FDsys will be accomplished in a trusted, secure environment built on the OAIS model, which takes a package-based approach to managing content. FDsys information packages are made of digital content and associated metadata, bound together by packaging information. The content includes various renditions, such as PDF or XML. In addition to descriptive metadata, the package also includes the preservation and technical metadata essential to the capability of preserving the content over time. This metadata also assists other systems to open and render the package intelligibly. Digital content is received from content originators as a Submission Information Package (SIP), managed and preserved as an Archival Information Package (AIP), made available for access as an Access Content Package (ACP), and delivered to end users as a Dissemination Information Package (DIP).

Document integrity is a fundamental goal of the FDsys preservation processes. The choice of renditions included in the Archival Information Package (AIP) is intended to support this goal. The top priority for preservation is the document's contents (e.g., text, graphics, etc.). To the extent that it is possible the original presentation "look and feel" should be preserved as well. AIPs consist of content information ("renditions") and associated preservation metadata needed to preserve the content over the long term. The preservation rendition in the AIP is a "normalized" rendition of the content intended to support future preservation processes. The preservation rendition will typically be an XML file, tagged to indicate where associated graphics should appear. These graphics will be included in the AIP. The AIP will also include other renditions, such as the as-submitted rendition from the content originator in its native format.

Preservation Processes

AIPs will be maintained in FDsys Preservation Storage. Within the preservation repository, GPO will assess the status of the archived packages and apply preservation processes to ensure that the archived content remains accessible and usable. In order of preference, the desired outcomes of the digital preservation processes are:

- Faithfully duplicated files, rendered using the original application.
- Files which faithfully reproduce content, behavior and presentation of the original, rendered using other software than the original application.
- Files which exactly convey the content but may alter behavior and/or presentation rendered using other software than the original application.

FDsys must be capable of supporting activities necessary to keep content accessible and usable. The preservation process employed in any given situation should be the least intrusive; i.e. that which alters the original AIP the least. A decision hierarchy to identify the most desirable

process to employ in a given situation is included in the *FDsys Requirements Document, version 3.2*. GPO's preservation strategies include:

- Refreshment (copying) of content to new media. Refreshment is the systematic transfer of stored digital information to newer, fresher media.

- Migration of data in formats or versions that are in danger of becoming or have become obsolete, to newer versions of that application or format. Migration is a process in which the underlying information is retained but older file formats and internal structures are replaced by newer.

- Emulation preserves the essential behaviors and attributes of digital objects by using current software to mimic the original environment.

- Hybrids of these approaches, or new approaches which have yet to emerge.

Secure Environment
FDsys' implementation of the OAIS reference model takes multiple approaches to ensuring a secure repository environment. First, in an extension of the OAIS model, in FDsys AIPs are maintained in a secure environment, and are not routinely "touched" by the system to satisfy user queries. Instead, the Access Content Package (ACP) is generated for this purpose, and from the ACP the query-specific Dissemination Information Package (DIP) to fulfill a user need is derived. Second, FDsys includes requirements for redundant, secure backup repositories, where additional copies of AIPs are maintained. Beyond the backup sites managed by GPO, other backup repositories may be operated by other institutions with which GPO has a contractual or partnership agreement, and FDsys requirements provide for the generation of exact copies of AIPs for this purpose.

FDsys and the TRAC Attributes
The FDsys requirements which bear on digital preservation have been reviewed against key attributes described in the Center for Research Libraries' 2007 *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist*. FDsys will meet the elements described in the technologies, technical infrastructure, and security sections of the TRAC; although some of the necessary capabilities for full compliance are enabled in later releases. GPO used the audit checklist to develop our requirements for the archive, and as system components are put into production, it is GPO's intention that FDsys comply with the intent of the TRAC attributes. These attributes include:

- Replication/Redundancy (more copies are safer)
  Redundant storage is an integral part of the FDsys design. FDsys preservation storage will consist of collections of AIPs with identical content located at multiple sites or in multiple instances of the system. This redundancy ensures preservation in the event of a disaster or other significant discontinuation of service at a single site. In addition to redundant storage, AIPs will be separate from working or production copies.

- Migration (move copies forward in time)

FDsys preservation strategies include refreshment, migration, and emulation. The preservation process employed in any given situation should be the least intrusive; i.e. that which alters the original AIP the least.

- Transparency (open source software)
  From the earliest phases, FDsys has been designed to preserve digital content free from dependence on specific hardware or software. The system will have the ability to transform content and metadata into formats that are free of proprietary restrictions. By design, FDsys AIPs will be self-describing so they may be opened and used in other systems and settings.

- Diversity (no single point of failure due to system "monoculture")
  FDsys will have the capability to produce content packages which fully replicate AIPs. In addition to having the preservation repository in a separate instance of the content management system, FDsys will have the capability to deliver content packages to other repositories.

- Audit (confirm data is preserved)
  FDsys will maintain content integrity and provide an audit trail of preservation processes. The system logs will record the results of the migration or refreshment process, including the ability to produce notification of an incomplete or unsuccessful migration or refreshment process.

- Economy (cost effective processes)
  Cost-effective digital preservation requires controlling ingest costs and metadata creation costs. FDsys will address both of these issues. Evaluation of content prior to ingest, performed by the system with human intervention as necessary, will ensure that only content in scope for GPO's dissemination programs will be ingested into FDsys archival storage and managed by preservation processes. In addition, FDsys will be capable of ingesting metadata that accompanies content and automatically creating technical, administrative, and preservation metadata. This system metadata will be enhanced and augmented by GPO metadata specialists.

Conclusion

In the years since entering the world of online dissemination, GPO has been building upon the foundation of the FDLP model which relies upon depository libraries as the trusted repositories for print content. It has expanded and extended the print delivery and preservation model into the digital world. GPO has progressed from delivering digital content derived from printing processes via GPO Access to authenticating that content using digital signature technologies. More recently GPO has successfully delivered authenticated content from a born digital source. At the system level, GPO has been recognized as an official NARA archival affiliate. With the development of the OAIS-standard based FDsys, GPO is poised to move to the next level and meet the guidelines for recognition as a trusted digital repository.

Glossary

United States Government Printing Office. Information Technology Glossary. Washington, DC: 2008. 3 July 2008. <http://www.fdlp.gov/repository/it-glossary-acronyms/index.html>.

References and Links

American Library Association, Association for Library Collections and Technical Services, Preservation and Reformatting Section, Working Group on Defining Digital Preservation, "Definitions of Digital Preservation." Washington, DC: 2007. 2 May 2008. <http://www.ala.org/ala/alcts/alcts.cfm>.

Center for Research Libraries. *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist, ver. 1.0*. Chicago, IL: Center for Research Libraries; Dublin, OH: OCLC Online Computer Library Center, Inc., 2007. 2 May 2008 <http://bibpurl.oclc.org/web/16712>.

Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS). Washington, DC: 2002. 29 Mar. 2006. <http://public.ccsds.org/publications/archive/650x0b1.pdf>.

Jantz, Ronald and Michael J. Giarlo. "Digital preservation: architecture and technology for trusted digital repositories." *D-Lib Magazine* 11/1 (June 2005) p. 3. <http://www.dlib.org/dlib/june05/jantz/06jantz.html>.

Lavoie, Brian. The Open Archival Information System Reference Model: Introductory Guide. Dublin, Ohio: OCLC Online Computer Library Center, Inc., 2004. 21 Mar. 2006. <http://www.dpconline.org/docs/lavoie_OAIS.pdf>.

Lavoie, Brian and Lorcan Dempsey. "Thirteen Ways of Looking at...Digital Preservation." *D-Lib Magazine* 10/7-8 (July/August 2004) 11 May 2006. <http://www.dlib.org//dlib/july04/lavoie/07lavoie.html>.

Research Libraries Group. *Trusted digital repositories: attributes and responsibilities.* Mountain View, CA: Research Libraries Group, 2002. p. 1 <http://www.rlg.org/legacy/longterm/repositories.pdf>.

United States Government Printing Office. A Strategic Vision for the 21st Century. Washington, DC: GPO, 2004.

United States Government Printing Office. Concept of Operations for the Future Digital System V2.0. 16 May 2005. 22 Mar. 2006 <http://www.gpo.gov/projects/pdfs/FDsys_ConOps_v2.0.pdf>.

United States Government Printing Office. Federal Digital System (FDsys) Requirements Document (RD) For Public Release, Revision 3.2 (RD-3.2), 4 December 2007. <http://www.gpo.gov/projects/pdfs/FDsys_RD_v3.2.pdf>.

United States Government Printing Office. Federal Digital System: System Design Document. R1C2 edition, May 2008. Unpublished.

United States Government Printing Office. Managing the FDLP Electronic Collection: A Policy and Planning Document. Washington, DC: GPO, 2008. 9 May 2008. <http://www.access.gpo.gov/su_docs/fdlp/pubs/ecplan.html>.

United States Government Printing Office. "Memorandum of Understanding between the Government Printing Office and the National Archives and Records Administration," 12 Aug. 2003. <http://www.gpoaccess.gov/about/naramemofinal.pdf>.

United States Government Printing Office. Study to Identify Measures Necessary for a Successful Transition to a More Electronic Federal Depository Library Program. Washington, DC: GPO, 1996.

Waters, Donald. "Good Archives Make Good Scholars: Reflections on Recent Steps Toward the Archiving of Digital Information." The State of Digital Preservation: An International Perspective Conference Proceedings. Washington: Council on Library and Information Resources, 2002. 11 May 2006. <http://www.clir.org/PUBS/reports/pub107/waters.html>