A Holistic Approach for Establishing Content Authenticity and Maintaining Content Integrity in a Large OAIS Repository

Lisa LaPlant, Kate Zwaard; United States Government Printing Office; Washington, D.C.

Abstract

This paper describes a holistic approach for establishing content authenticity and maintaining content integrity in a large OAIS repository. The U.S. Government Printing Office (GPO) is developing the Federal Digital System (FDsys), a digital archive to provide permanent public access to United States federal government publications. FDsys is being designed based on the OAIS reference model, which incorporates many archiving best practices and increases the chance for developing a successful archive that will protect and preserve content over time.

A cornerstone of any successful preservation system is a solid authentication strategy. GPO disseminates official United States federal government publications across the nation and around the world. Since 1861 users have looked to GPO as a trusted source for federal legal and regulatory information. This historic trust relationship increases the importance for GPO to establish content authenticity and maintain content integrity as it transitions from a publication-based print environment to a content-centric digital environment.

Defining, evaluating, and expressing content authenticity is essential for a successful authentication strategy. GPO defines authentic content as content that is verified by GPO to be complete and unaltered when compared to the version approved or published by a content originator such as Congress or a federal agency. Content will flow into FDsys in three distinct streams: deposited in digital format, converted from tangible media, and harvested from digital sources. Each avenue presents a unique challenge to evaluating authenticity and expressing chain of custody information (i.e., custodial history) for that content. Deposited content is submitted by a federal agency. In order to establish content authenticity at the time of submission GPO requires authentication of users including personal identity, corporate identity, and authorization to publish official federal government publications on behalf of their agency. Converted content presents a unique situation in that information must be recorded about the original hard copy publication and associated digitization processes. Establishing content authenticity for harvested content is often challenging because little information is available about the content creation process, and an expression of authenticity information may only indicate that the content was harvested from an official source.

Once content has been ingested into FDsys, the system maintains content integrity by providing assurance that content has not been changed in an unauthorized manner. GPO will use a multi-faceted strategy to provide this assurance including deploying a robust content management system, interrogating hashed content on a regular basis in order to detect changes, and applying digital signatures to PDF files that are delivered to public users.

Introduction

By law and tradition dating back to the late 1800s, the United States Government Printing Office (GPO) has had three essential missions:

- to provide the agencies and organizations which make up the three branches of the federal government with expert publishing and printing services, on a cost recovery basis, in order to avoid duplication and waste of government resources,
- to provide, in partnership with federal depository libraries, for nationwide community facilities for the perpetual, free and ready public access to the printed and electronic documents, and other information products, of the federal government, and
- to distribute, on a cost recovery basis, copies of printed and electronic documents and other government information products to the general public [1].

In order to continue to meet its missions, GPO's Federal Digital System (FDsys) is being developed to ingest, preserve, and provide access to authentic and official government publications from the legislative, executive, and judicial branches of the United States federal government. The system, which was launched as an internal proof of concept in August 2007 and is scheduled for an initial public release in late 2008, will automate many lifecycle processes for digital content and make it easier to deliver content in formats suited to customers' evolving needs [2].

The system design is based on the Reference Model for an Open Archival Information System (OAIS) developed by Consultative Committee on Space Data Systems (CCSDS) with broad input from other communities [3]. FDsys is designed around the concept of content packages, derived from the Warwick Framework, an approach in which discrete packages of metadata can be aggregated in conceptual containers [4]. FDsys packages, which exist for varying durations, contain metadata files, content files, and a binding Metadata Encoding and Transmission Standard (METS) XML file. A Submission Information Package (SIP) is a transitory package, which once ingested becomes an Archival Information Package (AIP). An Access Content Package (ACP) is also derived from the SIP. The ACP is designed to separate access renditions (e.g., HTML, PDF) from long term preservation renditions (e.g., XML). A Dissemination Information Package (DIP) is content and metadata delivered from the system [5].

Establishing Content Authenticity

For almost 150 years, the presence of the words "United States Government Printing Office" on a government publication assured the public that the content therein expressed government information as it was approved by a government author. Moreover, this assurance was strengthened by trust relationships that were established between all parties in the publication creation,

production, and dissemination process. Historically, a printing specialist from Congress, federal agency, or U.S. Court would contact a GPO customer service specialist to submit a publication to be printed. The resulting printed government publication would be made available to the public through the Federal Depository Library Program (FDLP), a system of more than 1,200 libraries in the United States that receive government publications from GPO and provide public access to those publications. The publication would also be made available through the GPO Sales Program where the public could purchase a copy from a GPO employee at a brick-and-mortar U.S. Government Bookstore.

In today's digital environment, GPO must assure users that publications available from GPO websites are as official and authentic as publications that have been printed and disseminated by GPO for nearly 150 years and that trust relationships exist between all parties in electronic transactions. For example, GPO must ensure that the regulatory information downloaded from the *Code of Federal Regulations* in FDsys is an authentic representation of the content submitted to GPO by the Federal Register Office within the National Archives and Records Administration (NARA).

In order to establish and express content authenticity in a digital environment, GPO must first define content authenticity, evaluate content authenticity during submission to the system, maintain content integrity while the content is in the system, and capture and display information that allows users to evaluate the authenticity and integrity of content delivered from FDsys.

Defining Content Authenticity

FDsys will contain authentic and official government publications. GPO defines *authentic content* as content that is verified by GPO to be complete and unaltered when compared to the version approved or published by a content originator such as Congress or a federal agency. *Official content* is defined as content that is approved by, contributed by, or harvested from an official source in accordance with accepted program specifications. A *Government publication* is a work of the United States government, regardless of form or format, which is created or compiled in whole or in part at government expense, or as required by law, except that which is required for official use only, is for strictly operational or administrative purposes having no public interest or educational value, or is classified for reasons of national security [6].

Evaluating Content Authenticity

FDsys content will be deposited in digital format, converted from tangible media, and harvested from digital sources. Each avenue presents a unique challenge to evaluating and establishing content authenticity. Deposited content, which will include all common text, graphical, audio, and video formats used in government publishing, is intentionally submitted to GPO by U.S. federal content creators or their authorized designees [6].

In order to record the chain of custody (i.e., custodial history) and establish content authenticity at the time of content submission, GPO will authenticate users depositing content into FDsys based on personal identity, corporate identity, and authority to publish official federal government publications.

While content authenticity can be readily established for deposited content, the authenticity of converted content is not as apparent. Converted content is digital content created from a tangible product, typically by scanning legacy print publications. An evaluation of the authenticity of converted content must take into account the source of the original tangible content and any digitization processes applied to content to produce digital files for submission to FDsys and ingest to the preservation repository.

Harvested content includes government publications that are gathered from federal agency websites or other sources in accordance with accepted program specifications. Establishing custodial history and evaluating content authenticity for harvested content is often challenging because little information is available about the content creation process. An expression of authenticity information may only indicate that the content was harvested from an official source.

Maintaining Content Integrity

Title 44 of the *United States Code* stipulates that public access to official government publications disseminated through the FDLP must be maintained permanently in regional depository libraries [7]. Since online publications are not physically distributed to depository libraries for retention, GPO has assumed responsibility for the provision of permanent access to federal government publications residing on GPO's servers. GPO's permanent public access commitment is also met by bringing agency-disseminated electronic publications under the purview of GPO and incorporating them into a digital preservation repository that maintains content integrity. Content integrity refers to providing an assurance that content has not been altered or destroyed in an unauthorized manner that information is recorded about any authorized changes.

Maintaining content integrity for publications that fall within scope of the FDLP is paramount, because in many cases GPO is responsible for maintaining the system of record for official federal government publications. For example, NARA recognizes GPO as an official archival affiliate for electronic federal government publications on the GPO Access website. Signed in 2003, the NARA-GPO agreement provides for the permanent preservation and access to the *Congressional Record*, the *Federal Register*, the *Code of Federal Regulations*, and other electronic federal government publications. Because of this and other commitments, FDsys will ingest, preserve, and provide public access to electronic federal government publications that are in scope for the FDLP, including publications currently available on the GPO Access website [8].

FDsys system functions comprise distinct subsystems that support 1) content submission and authorized user access, 2) ingest into the preservation repository and preservation processes, and 3) public access.

Submission Processing

Content and metadata are submitted to FDsys prior to ingest into the preservation repository (i.e., archive). When a user first uploads a content file to FDsys, it is inspected for malicious code. The SHA-256 cryptographic hash function is then applied to content files and a message digest (i.e., hash value) is computed for each file. Content and metadata may be added or removed until the user approves it for publication and submits the SIP to FDsys. The submission action by the user triggers background processes including

- determining if content is a new version of existing content,
- parsing metadata from the content files,
- breaking up large files as necessary for ease of access, and
- performing SIP validation including verifying that metadata files comply with required XML schema and include mandatory metadata elements.

Ingest Processing

Upon completion of all tasks associated with submission processing, an XML-based preservation rendition is created. The preservation rendition and all submitted renditions are added to the content package for ingest into the archive. A digital time-stamp is applied to the preservation rendition and an additional hash value including the time-stamp is calculated and stored in metadata. Hash values are stored in packaging metadata, which is not editable by any system user. The package is then ingested into the preservation repository and moved to archival storage.

Preservation Processing and Repository

The preservation repository contains AIPs which are stored separately from a working copy of the package, the ACP. In addition to separate physical storage, the preservation repository relies on separate processors from those that manage SIPs and ACPs. While both the preservation repository and the access repository use a commercial off the shelf (COTS) content management system, the preservation repository is a separate instance of the COTS product. Processes synchronize metadata between the two repositories as necessary, and hash values are used to evaluate content integrity within a repository and between corresponding content files in separate repositories. The FDsys security model, which is based on user roles and groups, strictly enforces policy specifying that only preservation specialists can modify content packages in the archive. Content integrity is also maintained when preservation processes, such as migration, are performed on content. Furthermore, FDsys replication and back-up functions for all subsystems contribute to GPO's ability to maintain content integrity.

Access Processing and Repositories

The access repository contains the ACP, which is a working copy of the content package. The ACP is created from the SIP after ingest and includes all submitted files plus a copy of the preservation rendition. The ACP also contains access derivative renditions (e.g., HTML, PDF) and renditions that are part of the print production process (e.g., Postscript, SGML, Quark). Authorized users have the ability to access ACPs, add renditions to ACPs, delete renditions from ACPs, and modify metadata in the ACP based on privileges associated with their role and group. Changes to ACP descriptive and technical metadata are automatically propagated to the AIP where appropriate, while additions and deletions of access renditions are not propagated to the AIP. In fulfillment of Title 44 of the *United States Code* and GPO's missions, access renditions from content packages that are in scope for the FDLP and associated content metadata containing descriptive, provenance, and fixity information are added to the public access repository. ACPs in the public access repository are indexed by a COTS enterprise search engine and available for public users to search, download, and print as DIPs. FDsys processes ensure that corresponding content and metadata in the access repository and public access repository remain synchronized.

Digitally Signed PDF Files

Electronic publications pose special challenges to content integrity and authenticity because they can be easily altered or copied, leading to multiple non-identical files that can be used for unauthorized or illegitimate purposes. To further ensure the authenticity and integrity FDsys content, GPO uses a digital certificate to apply a digital signature to PDF files in the access repositories. In February 2008, GPO worked with Office of Management and Budget (OMB) to make digitally signed PDF files for the fiscal year 2009 *Budget of the United States Government* available via the GPO Access website, and in March 2008, GPO began digitally signing PDF files of public and private laws for the 110th Congress, which are also available via the GPO Access website.

The digital signature process establishes GPO as a trusted information provider and provides assurance that a PDF file has not been altered since it was disseminated by GPO. This assurance is important because the PDF rendition corresponds directly to the printed publication which continues to be the preferred format for legal citation. Visible digital signatures on PDF files serve the same purpose as handwritten signatures or traditional wax seals on printed publications.

A digital signature on a PDF file ensures content integrity and authenticity at no additional cost to public users. Upon opening a digitally signed PDF file in the widely available free Adobe Reader, public users are alerted if the file has been altered in an unauthorized manner and can choose to view the unaltered version. Furthermore, users are able to validate the certificate that was used by GPO to apply a digital signature to the file. This is one of many processes to express the authenticity and integrity of content available from FDsys.

Expressing Authenticity

An archive must communicate authenticity and trust to users on two levels – at the content level and the system level. In FDsys, content authenticity after submission is documented and expressed to users through provenance metadata. In order for that information to be useful, however, users must trust the archive as an entity. FDsys communicates trustworthiness at the system level through transparency in design and by following the Trustworthy Repositories Audit & Certification: Criteria and Checklist.

Expressing Content Authenticity and Integrity

An AIP consists of content information, Preservation Description Information (PDI), and other metadata. PDI, as described by the OAIS reference model, is information necessary to manage the preservation of the targeted content by recording the following four types of information:

- *Provenance* is the information that documents the history of the content information. It records the source, the changes that have occurred since it was created or acquired, and who has had custody of it. This gives users some assurance as to the likely reliability of the content information.
- *Context* documents the relationships of the content to the environment (why it was created and how it relates to other content).

- *Reference* information provides ways to refer to the content (e.g., ISBN, DOI).
- *Fixity* is information used to ensure the content object has not been altered in an undocumented manner (e.g., hash value).

Of the four types of PDI, provenance, context, and fixity play a role in communicating authenticity and integrity to users.

Providing fixity information to users helps assure them of the content's integrity. If the integrity of an object is compromised, it cannot be authentic.

Provenance, from an archival perspective is a special type of context information. Context information is described by OAIS as "information that documents the relationships of the Content Information to its environment. This includes why the Content Information was created and how it relates to other Content Information objects existing elsewhere" [3]. Context is related to the archival principle *respect des fonds,* which means that the archive should maintain a corpus of records together as a unit [9]. (The alternative would be to organize each record separately, as for example, with library books.) FDsys maintains context provenance by creating references to related objects in metadata.

Provenance information as defined by OAIS is closer to the concept of custodial history, which is "the chain of ownership of the materials being described, before they reached the immediate source of acquisition. Both physical possession and intellectual ownership can be described, providing details of changes of ownership and/or custody that may be significant in terms of authority, integrity, and interpretation" [10]. This definition can be extended before acquisition to the date of creation so that provenance metadata from another trusted repository can be ingested into FDsys along with the acquired object, giving a more comprehensive view into its history.

Provenance information helps assure users and archive administrators that significant events in the content lifecycle are documented and reveals whether and how much significant properties of an object have been altered. Objects in the system are authentic so far as they are unaltered compared to the version that was first acquired by the archive, with respect to their significant properties. Significant properties are the attributes of an object that affect its "quality, usability, rendering, and behavior" [11]. The tolerance for type and degree of change to significant properties before an object can be considered inauthentic varies by cultural heritage institution and designated community. For example, a rendition with a slight shift in color could be considered authentic by FDsys, but may be rejected by an art museum. In this example content integrity was not violated; the change to content was authorized but resulted in a possibly inauthentic object. Expressing chain of custody in metadata allows information professionals and public users to evaluate the authenticity of an object in the archive, while allowing standards to change according to the needs of the community.

FDsys is using the PREMIS data dictionary for guidance in collecting, organizing, and displaying provenance information to users. Some of the events for which FDsys will create an event entity (adapted from the data dictionary [12]), are as follows

- unpacking – extracting an object from a compression or packing file
- compression – encoding data to save storage space or transmission time
- virus check – scanning a file for malicious code.
- ACP creation – packaging of an ACP
- ingestion – adding AIPs to the preservation repository
- message digest calculation – creating a message digest
- fixity check – verifying that an object has not been changed in a given period
- digital signature assignment – digitally signing an object
- digital signature validation – determining that a decrypted digital signature matches an expected value
- public access restriction – removing a package from the access repository
- deletion – removing a digital object from repository
- deaccession – removing all of the content from a package from the inventory of a repository
- rendition creation – transforming an object to create a rendition in an additional file format or fidelity
- normalization – transforming an object to create a rendition more conducive to preservation
- migration – transforming an object to create a rendition in a more contemporary file format
- refreshment – moving an object from one storage system to another

Each event will be linked to a named agent in the system. All named users who can act on objects can be an agent linked to an event. Agents also include significant system components, including version numbers, so that if there is a faulty procedure, preservation specialists can track the results and contain any damage.

Communicating Trustworthiness

To address how repositories could communicate trustworthiness to their designated community and content partners, the Digital Repository Certification Project was created by the Research Libraries Group and the National Archives and Records Administration in 2003. Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC) is a report published by the project to help repositories objectively probe their claims of trustworthiness. The scope of the TRAC checklist is far more comprehensive than the authenticity and integrity of content information, addressing, for example, the archive's mission, the organization's policy framework and financial stability, and the repository's preservation strategies. The checklist is also a tool to communicate an organization's willingness and ability to maintain objects so users can trust their authenticity.

Below are some of the criteria outlined in TRAC that concern authenticity and integrity.

- B1.3 Repository has mechanisms to authenticate the source of all materials.
- B2.8 Repository acquires preservation metadata (i.e., PDI) for its associated Content Information.
- B4.4 Repository actively monitors integrity of archival objects.
- B6.10 Repository enables the dissemination of authentic copies of the original or objects traceable to originals.
- C1.5 Repository has effective mechanisms to detect bit corruption or loss.
- C3.3 Repository staff have delineated roles, responsibilities, and authorizations related to implementing changes within the system [13].

Providing evidence that FDsys has met these and other criteria allows GPO to demonstrate to the preservation community that it is using best practices for establishing the authenticity of content and maintaining integrity. Until the archive is certified, however, GPO communicates the trustworthiness of the repository through transparency in the software development lifecycle by sharing important deliverables at each phase, such as the Concept of Operations and Requirements Document.

Related Work and Next Steps

While this paper outlines an initial approach for establishing content authenticity and maintaining content integrity in a large OAIS repository, opportunities exist to explore various ways to implement and enhance this approach. Areas of interest include

- costs and benefits of using an external time-stamp authority,
- ways to enhance the integrity of hash values, such as a redundant hash value list,
- relative merits of a hash-signing approach where a hash value is digitally signed versus a hash-linking approach where a hash value is combined with other hash values received during the same time period,
- technologies for expressing the authenticity of content below the publication level of granularity,
- implementation strategies for expressing the integrity of content files in file formats other than PDF that are delivered to public users,
- ways to authenticate users who are submitting content, and
- methods for maintaining content authenticity and integrity during preservation processes.

To this end, FDsys will continue to monitor and evaluate other comparable platforms and systems including DSpace, an open-source repository platform; HP's Digital Media Platform, a service-oriented architecture for content processing and storage [14]; the British Library's Digital Object Management system, a system to enable the United Kingdom to preserve and use its digital intellectual heritage [15]; and the National Library of the Netherlands' (Koninklijke Bibliotheek) Digital Information and Archiving System (DAIS), the library's digital deposit repository [16].

References

[1] United States Government Printing Office, "A Strategic Vision for the 21st Century", (2004).

[2] GPO's Federal Digital System (FDsys). <http://www.gpo.gov/projects/fdsys.htm>.

[3] Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System (OAIS)," (2002).

[4] Carl Lagoze, "The Warwick Framework: A Container Architecture for Diverse Sets of Metadata." Cornell University – D-Lib Magazine. (July/August 1996).

[5] Gil Baldwin, Matthew Landgraf, Kate Zwaard, John Faure. "Content Packaging Approach for a Large OAIS Repository." In Proceedings of Archiving 2007, May 2007, pp 44-47.

[6] Authentication: Definitions and Acronyms. <http://www.gpoaccess.gov/authentication/definitions.html>.

[7] United States Congress. "Depository Library Program" Title 44 U.S. Code. Chapter 19, 2000 edition.

[8] GPO Access. <http://www.gpoaccess.gov>.

[9] Michael J. Fox, Peter L. Wilkerson. Archivist's Primer. (Getty Information Institute, 1998). <http://www.getty.edu/research/conducting_research/standards/introarchives/>.

[10] "Encoded Archival Description Tag Library" (Society of American Archivists, 2002). <http://www.loc.gov/ead/tglib/index.html>.

[11] Margaret Hedstrom, Christopher Lee. Significant Properties of Digital Objects: Definitions, Applications, Implications. Proceedings of the DLM-Forum, pg. 218. (2002). <http://www.ils.unc.edu/callee/sigprops_dlm2002.pdf>.

[12] OCLC and RLG, "Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group" (2005).

[13] CRL, NARA, "Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist." <http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=91>.

[14] Stuart Haber and Pandurang Kamat. "Content Integrity Service for Long-Term Digital Archives." In Proceedings of Archiving 2006, May 2006, pp 159-164.

[15] Adam Farquhar, Sean Martin, Richard Boulderstone, Vince Dooher, Richard Masters, and Carl Wilson, The British Library (UK). "Design for the Long Term: Authenticity and Object Representation." In Proceedings of Archiving 2005, pp 104-108.

[16] Raymond J. van Diesen, Titia van der Werf-Davelaar, "Authenticity in a Digital Environment: Long Term Preservation Study Report Series Number 2." (IBM Netherlands, Koninklijke Bibliotheek 2002).

Author Biography

Lisa LaPlant has worked for the U.S. Government Printing Office since 2001 and is currently a Lead Program Planner in the Program Management Office, which is responsible for planning and implementing FDsys. Her areas of expertise include user interfaces, search, and content authentication. Prior to working on the FDsys program, she worked on various aspects of the GPO Access website, including web design and outreach activities. She received a B.A. in Media Arts and Design from James Madison University in 2000.

Kate Zwaard is a Lead Program Planner in the Program Management Office for FDsys. Her areas of expertise include digital preservation, metadata management, and content authentication. Before she joined the FDsys team, Zwaard was responsible for data analysis and statistics for the electronic dissemination initiatives at GPO. She presented at Archiving 2007 on a content packaging approach for a large OAIS repository. She was graduated from the University of Maryland in 2002 with a double major in political science and journalism with specialties in public opinion and statistics.