U.S. GOVERNMENT PRINTING OFFICE
Keeping America Informed | www.gpo.gov

Authenticity of Electronic Federal Government Publications

U.S. Government Printing Office
Washington, D.C.
06/13/2011

Revision History

Revision	Date	Description
1.0	June 13, 2011	First published version

Table of Contents

I. Introduction

Since 1861 users have looked to the U.S. Government Printing Office (GPO) as a trusted source for Federal government information. The presence of the words “United States Government Printing Office” on a printed publication assures the public that the content inside expresses information as it was approved by a government author. This assurance is strengthened by trust relationships established between all parties in the creation, production, and publication process. In the traditional printing environment, a printing specialist from Congress, Federal agency, or U.S. Court contacts a GPO customer service specialist to submit a publication to be printed. The resulting printed government publication is made available to the public from GPO’s online bookstore or through the Federal Depository Library Program (FDLP), a system of more than 1,200 libraries in the United States.

GPO’s mission has not changed, but the adoption of digital technology has changed the ways its products are created, managed, and delivered to users. Verifying the authenticity of electronic documents poses a special challenge because electronic files are easily altered, which could lead to unauthorized modifications of government content used in illegitimate ways. GPO must assure users that electronic publications available from GPO websites are as authentic as the publications that have been printed and disseminated by GPO for nearly 150 years.

At its essence, GPO’s role in the authentication of Federal government publications has not changed; regardless of format, GPO strives to provide tools and evidence to allow users to determine the authenticity of content. For 150 years of ink-on-paper, the agency provided this evidence in the form of imprimatur on publications and the availability of publications from trusted sources (e.g., Federal depository libraries and GPO bookstores).

GPO defines **authentic content** as the complete and unaltered representation approved or published by the content originator or an authorized derivative with a trusted chain of custody to that representation. This definition creates a model for assuring the authenticity of electronic government information, regardless of changes in technology.

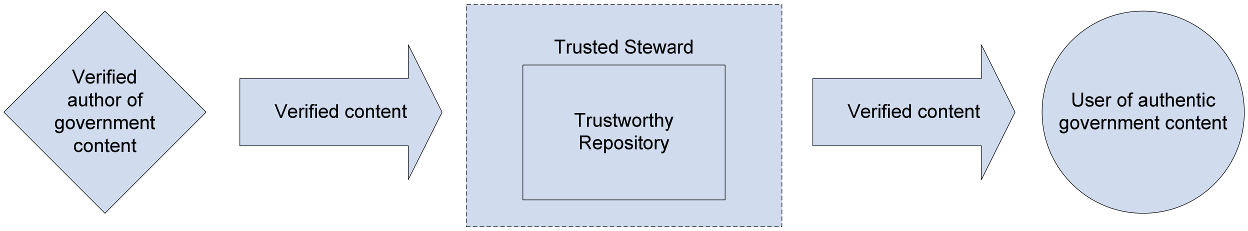
Verified author of government content
Verified content
Trusted Steward
Trustworthy Repository
Verified content
User of authentic government content

Figure 1: Trustworthy chain of custody assures users of the authenticity of information

In order to be satisfied that an item is authentic, users must be sure that 1.) they can trust the source of the content, and 2.) unauthorized alterations to content have not occurred (i.e., content integrity is maintained).

The remainder of this document describes the tools and evidence that GPO provides to users to help them verify these two attributes. In the case of the first attribute, trustworthy source, GPO provides evidence that the electronic information it maintains is from a trustworthy repository and the history of each item in the repository can be documented. In the case of the second attribute, content integrity, GPO provides content integrity tools such as digitally signed PDF files and cryptographic hash values.

II. Trustworthy Source

As discussed in the introduction, for most of GPO's history, users relied on knowledge of trust relationships between parties in the publication chain to be sure they were using official government information. Since the advent of word processing, the amount of information being created has exploded, and relying on these same interpersonal trust relationships is not scalable. How can users determine that the source of content is trustworthy and that the representation of content they are looking at (such as an EPA Report in HTML) is an authorized derivative of what was submitted to GPO (for example, the same EPA Report in PDF)?

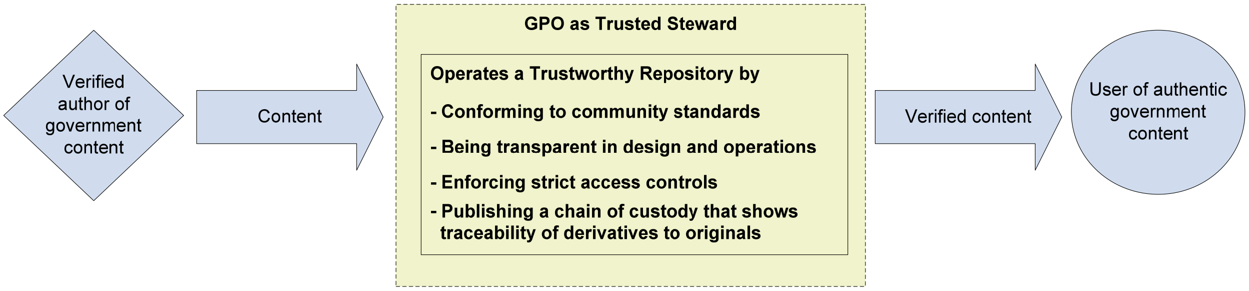
Verified author of government content
Content
GPO as Trusted Steward
Operates a Trustworthy Repository by
- Conforming to community standards
- Being transparent in design and operations
- Enforcing strict access controls
- Publishing a chain of custody that shows traceability of derivatives to originals
Verified content
User of authentic government content

Figure 2: GPO as a trusted steward operates trustworthy repository that documents the chain of custody of content and maintains its integrity

A. GPO as a Trustworthy Steward of Information

For 150 years, official Federal government information has been printed and distributed to the public through GPO. In 1993, with the advent of *GPO Access*, GPO became a repository and disseminator of official, no-fee electronic publications from all three branches of the Federal Government.

In response to the growing need for tools to preserve, manage, and provide access to Federal government content, GPO developed Federal Digital System (FDsys) (www.fdsys.gov). As the collection grows, FDsys will provide the American people a one-stop site to access authentic, published government information. The system is envisioned as a comprehensive, systematic, and dynamic means for preserving any type of digital content, independent of specific hardware or software. Six months after launch, FDsys was named one of the top government websites by *Government Computer News*.

To address how repositories could communicate trustworthiness to their community and content partners, the Digital Repository Certification Project published Trustworthy Repositories Audit & Certification: Criteria and Checklist, often referred to as TRAC (RLG-NARA Task Force on Digital Repository Certification, 2007), a set of attributes the preservation community has agreed is important for digital repositories. 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 digital objects so that users can trust their authenticity. FDsys was created with these criteria in mind and conforms to all technical requirements in the TRAC checklist.

Providing evidence that FDsys has met these and other criteria allows GPO to demonstrate to the preservation and user community the use of best practices for establishing the authenticity and maintaining the integrity of content. GPO is working towards certification by an auditor as a trustworthy repository and will continue to foster trust through transparency in the software development lifecycle by presenting at conferences and publishing documentation on GPO's website (http://www.gpo.gov/fdsysinfo/aboutfdsys.htm). Examples of documentation made available include the Concept of Operations, Requirements Document, and design artifacts.

To meet the needs of the user community and the requirements set forth by TRAC, GPO maintains the integrity of content in the repository by

1. Restricting access to content in the repository, and
2. Ensuring content in the repository has not been added, altered or removed without authorization.

1. Restrictions on Access to Content

FDsys security strictly controls access to electronic content in the repository. No user has access to alter content files. Users can submit new content and change the descriptive metadata about content, but it is not possible to open a file in the repository, change a sentence, and then save it. The content a user sees on the public GPO website is a copy of content taken from the preservation master content in the archive. FDsys processes ensure that corresponding content and metadata in the archive and access repository remains synchronized, and the archival copy of content can be used to regenerate the access copy at any time. The FDsys security model, which is based on user roles and groups, strictly enforces policy specifying that only preservation specialists can access content packages in the archive.

2. Ensuring Content Has Not Been Altered

FDsys periodically checks content in the system for

1.) Corruption or changes to content,
2.) The presence of an unauthorized content file or package, or
3.) The absence of an expected content file or package.

This polling is accomplished by calculating a SHA-256 cryptographic hash value from the content and comparing it with the value recorded at time of ingest. If the system detects any changed, missing, or unauthorized content, GPO system administrators will be notified and can take action, such as initiating a security assessment and restoration of content from backups.

This section on maintaining content integrity has described

- GPO's processes for preventing changes to content between receipt from publisher and delivery to users, and
- GPO's tools for detecting changes to content between receipt and delivery.

The next section on chain of custody details how users can tell that the representation of content they are viewing was published through the same kind of trust relationships that exist in the traditional hardcopy production processes, and that it is functionally the same as the representation submitted by the content originator.

B. Chain of Custody

Chain of custody, or provenance,[1] for digital libraries and archives is the information that documents the history of the content information. Chain of custody records the source, the changes that have occurred

[1] Provenance is a French term that means "to come from," and it is used in a wide range of fields to mean information that allows users to make a determination that the item they are inspecting is, in fact, what it purports to be. For example, the provenance of the Jasper John's painting, *Target with Plaster Casts,* would list all of its owners from the time it left the property of Mr. Johns to its current home in New York (in other words, the painting's chain of custody). If it were to be sold, buyers could inspect the evidence for its provenance to decide whether it is an authentic Jasper Johns painting.

since it was created or acquired, and who has had custody of the content information. This gives users some assurance as to the likely reliability of the content information.

GPO collects and provides to users information about each significant event in the lifecycle of content in the PREMIS.xml metadata file. The event information includes what occurred, who triggered the event,[2] what specific files the event affected, and the date and time of the event. FDsys uses the PREMIS data dictionary (PREMIS Editorial Committee, 2008) for guidance in collecting, organizing, and displaying provenance information to users. For a partial list of the events GPO is recording, please see Appendix 2: Events Captured in Metadata.

Chain of custody information 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 digital object, giving a more comprehensive view of its history.

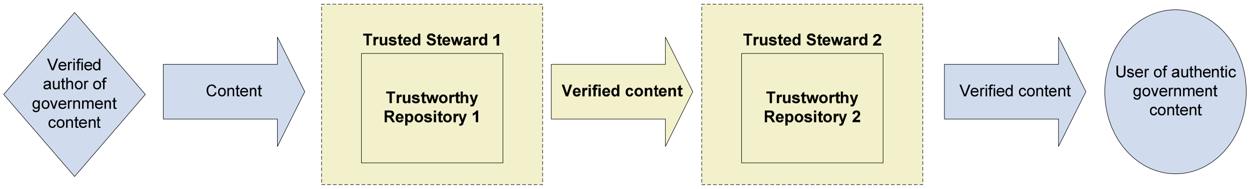

Figure 3: An extension of the model in which GPO receives content from another trusted source

As GPO’s digital collection grows, preservation metadata can be used to track down any potential problem areas. For example, if there is a transform between one file format to another that was discovered to be faulty, metadata about the agent allows system administrators to uncover every object created using the faulty transform. Similarly, if a user is found to have submitted unauthorized material, administrators can isolate all content that user acted on and take necessary action.

Chain of custody is especially important in helping users evaluate the authenticity of derivatives of content. These derivatives are created for two reasons: to perform preservation processes and to provide an alternative representation of content.

As software evolves, some file formats will fall out of use and be difficult to render intelligibly. One strategy GPO uses to ensure the continued usability of content is to move the data to a more current file format. For example, *The Economic Report of the President* has an appendix that provides additional information in the form of spreadsheets, some of which were given to GPO in Word Perfect, a format most users are no longer able to open. As that content was migrated from *GPO Access* to FDsys, the Word Perfect files were kept in the archive, and GPO created more user-friendly derivatives through a preservation process.

[2] When each event is recorded in the preservation metadata, the agent responsible for triggering or enacting the event is also recorded. Agents are either a named system user or a software component, including version number.

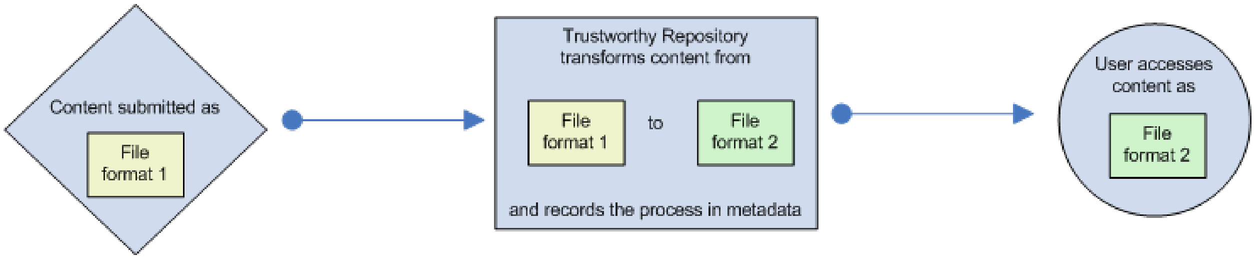

Figure 4: Content is transformed from one representation to another to preserve the content or to make the content easier to access

The format in which content is submitted to GPO is often more suitable to the creation of content than for electronic access. GPO saves the submitted files in the archive but will provide access to the content in a more common format.[3]

Representations of content created by GPO, or other trusted parties, where the chain of custody can be traced back to the originally submitted representation are considered by GPO to be authentic Federal government content. This does not always mean that the content is identical to the original, but it does mean that, to the extent possible, the "essence," or significant properties, of the content has been preserved.[4]

III. Communicating Content Integrity

Maintaining content integrity means ensuring that content has not been changed or destroyed without authorization. This is especially important for electronic information, since minor (but critical) changes to content are difficult for a user to detect, and if the integrity of an object is compromised, it cannot be authentic. The type of information that digital repositories and users employ to determine that content has not changed is called **fixity** information.

GPO provides the following tools to assure users that the integrity of the content has not been compromised:

A.) cryptographic hash values in metadata, and

B.) digital signatures on PDF files.

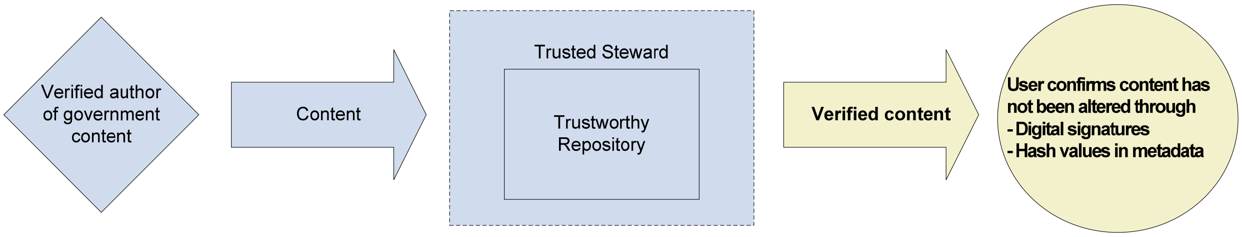
Verified author of government content
Content
Trusted Steward
Trustworthy Repository
Verified content
User confirms content has not been altered through
- Digital signatures
- Hash values in metadata

Figure 5: GPO provides tools and evidence for users to determine that content has not been altered since it was received by GPO

[3] For example, when a brochure is submitted in a file format created from a desktop publishing program like InDesign, GPO will provide that content to users in a more commonly used format, such as PDF or HTML. In another example, if a 1972 issue of the *Congressional Record* is scanned and saved as an image file, GPO will run the image file through optical character recognition software to create a text file from the picture of words in the image and provide the resulting image to users.

[4] The creation of a derivative of a representation, by its very nature, will entail some changes from the native representation. When GPO creates or accepts a derivative (for access or preservation purposes), it uses quality control measures to minimize changes to the "essence of the document," or its significant properties. Significant properties are the attributes of an object that affect its quality, usability, rendering, and behavior.

A. Cryptographic Hash Values

Cryptographic hash values are provided to users for every publicly-accessible file on FDsys in the PREMIS.xml metadata file. Users can search for content on FDsys and, using the hash values and publicly-available tools, check for themselves that the content has not been altered.

B. Digital Signatures on PDFs

GPO first defined its strategy for content authentication in a 2005 white paper that described the agency's need to develop policies and create systems addressing the authentication of electronic government publications. Since PDF is the most common file type used for public access in FDsys, GPO began applying digital signatures to PDF documents in 2008, starting with the President's Fiscal Year 2009 Budget. These digital signatures provide a visible assurance to users that a document has not been altered since it was downloaded from GPO's websites. GPO is continuing to digitally sign additional publications, based on the approval of the content originator.

PDF documents digitally signed by GPO will display a blue ribbon on the first page if the document has not been modified since the signature was applied. Signature properties are also available for users who require further information on the certificate.

The chief advantages of digital signatures on content files are

- A user who receives a PDF published by GPO from a secondary source (e.g., an email from a colleague or a private website) can use the digital signature to assure themselves that the content has not been altered, and
- The technology allows users to verify GPO as the source of the material.

For more information about digital signatures, see Appendix 1: How Digital Signatures Work.

IV. Investigation into Other Requirements for Content Authentication

GPO recognizes that as technology changes and the field of digital content authentication develops, requirements for policies and technologies will change. GPO strives to evolve and will continue to be flexible and adapt its practices to meet the needs of its users and content authentication best practices. To that end, GPO constantly monitors the user community and other practitioners for new requirements or models to follow. Below are two examples of capabilities GPO is currently investigating.

A. Verification of Transmission from Government Author to GPO

A key link in the chain of custody is the transmission from government author to GPO. In most cases, verification of this transmission is done manually, through GPO's trust relationships with its agency customers. But just as transmission between FDsys and the end user is verifiable through digital signatures and hash values, similar technology can be applied in the front end of the process to ensure that GPO is, in fact, receiving what the author thinks it has sent. GPO is already employing some of these tools. For example, when GPO receives the United States Budget, it validates a digital signature associated with the transmitted files.

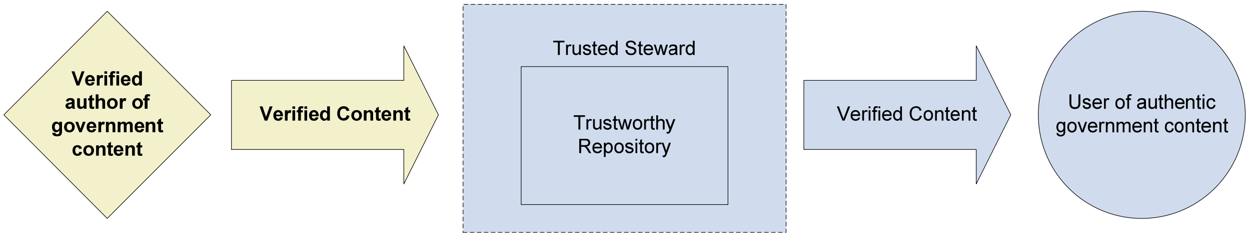
Verified author of government content
Verified Content
Trusted Steward
Trustworthy Repository
Verified Content
User of authentic government content

Figure 6: To complete the chain of custody, all content files would be matched bit-by-bit before and after transmission

B. Bulk Integrity Verification of XML Content

GPO stakeholders have expressed a strong interest for the agency to provide Federal government publications in XML for bulk download so it can be reused and repurposed by other information professionals. GPO is working hard to meet this need. In GPO's bulk repository and through the Data.gov portal, users can download large data sets of content in XML for reuse and data analysis; the most recent example of which is the use of GPO-supplied XML in the Federal Register 2.0 website.

The tools that GPO provides for individual users to confirm the authenticity of content (cryptographic hash values and chain of custody in metadata) are available for XML, but there may be a need for methods that are more easily scalable to the automated authentication of large sets of data. GPO is investigating the use of new technologies to enable bulk content integrity assurance of XML files. GPO does not intend for this technology to supplant or replace the other tools it uses for content authentication, but rather as an enhancement to the GPO content authentication program with a supplemental channel for high-volume, automated bulk data interfaces that prefer and require XML data feeds.

The publication of the cryptographic hash values in the PREMIS metadata file, and the way FDsys structures its public URLs, makes it possible for machines to crawl and use this information to determine content integrity in bulk. GPO is currently searching for users who are interested in bulk content integrity so the agency can better understand whether this method would meet their requirements. GPO recognizes the importance of ensuring that any content integrity verification method for XML content, such as digital signatures, should be structured so as not to interfere with data re-use or re-purposing. GPO is also committed to the principle of employing open, internationally recognized standards whenever possible.

C. User Authentication of Content Smaller than a Traditional Granule

GPO's model for a content package is roughly equivalent to a bound volume. Recognizing that users are usually looking for more discrete pieces of information, in many cases, GPO will break up submitted files into smaller logical units. For example, one issue of the *Federal Register* is packaged together, but users can search and download files at the article level. The tools used to determine the authenticity of content, as described above, are available for granules at this level, including digital signatures on the PDF files of granules.

GPO is investigating interest in and feasibility of helping users determine the authenticity of units of content even smaller than these granules. For example, users may want to know whether a single line in an appropriation bill is authentic, which means determining whether, 1) the line appears in the bill exactly as written, and 2) the line appears in the appropriation bill they thought it did, and not in last year's bill, for example. Based on feedback from the user community, GPO is gathering requirements for and looking into the feasibility of providing a service that would

- Allow users to specify what text they are interested in determining the authenticity of, rather than restricting to pre-formed granules,
- Allow users to easily and quickly determine which GPO-published documents contain a given set of text (or bitstream, which could provide extensibility to audio or video in the future),
- Provide a method for users to determine which document is of interest, if multiple documents in the repository contain the target text, and
- Produce a digitally signed document in "real-time" that indicates that user-specified text is contained verbatim in a given GPO-published file.

GPO plans to conduct feasibility studies of this concept, and, if it is determined to be technically and practically feasible, hopes to provide a prototype or proof of concept of this functionality at some point in the future to help facilitate discussion and investigation into this area of authentication.

D. Digital Signature Display on Mobile Devices

More and more users are using mobile devices for their media consumption, and GPO is looking to fulfill that need by providing products and services optimized for viewing on those platforms. Ideally the user experience on a mobile device, such as a smart phone, would be functionally identical to the experience on a personal computer. One of the current gaps between the two platforms is the inability to view digital signatures and certificate information when using the PDF reader provided on most mobile devices. GPO continues to talk with industry partners about including that capability in their development roadmaps.

Conclusions

GPO is a trusted steward of authentic Federal government content, and maintains control over the content lifecycle in a preservation system. GPO's policies and technologies are developed around a user-centric approach to content authentication, where the agency provides a suite of tools to help users make determinations about the authenticity of a particular piece of content. As the field of content authenticity develops, technology changes, and user requirements are identified, GPO's policies and technologies will continue to evolve. To this end, GPO, in partnership with the Library of Congress, is leading a Federal working group for content authentication to increase collaboration and standardization across the Federal government. The working group will be advised by a panel of experts from industry, academia, and government. GPO hopes to remain a leader in the field of content authentication by listening carefully to its stakeholders and driving innovation. Your input is appreciated, and any comments or questions can be directed to authentication@gpo.gov.

For general information about GPO's authentication program, see the 2011 authentication summary document, *Overview of GPO's Authentication Program* http://www.gpo.gov/pdfs/authentication/authenticationoverview.pdf.

Resources

Audrey Novak, Yale University Library. (2006). *Fixity Checks: Checksums, Message Digests and Digital Signatures.* <http://www.library.yale.edu/iac/DPC/AN_DPC_FixityChecksFinal11.pdf>.

RLG-NARA Task Force on Digital Repository Certification. (2007). *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist.* <http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf>.

Division of Administrative Rules. (2011). “Authentication.” *Utah State Digest.* <http://www.rules.utah.gov/publicat/digest.htm#Authentication>.

Farquhar, A., Martin, S., Boulderstone, R., Dooher, V., Masters, R., Wilson, C. The British Library (UK). (2005). “Design for the Long Term: Authenticity and Object Representation.” *Archiving 2005*.

Haber, S., Kamat, P. (2006).*Content Integrity Service for Long-Term Digital Archives*. *Archiving 2006.*

LaPlant, L., Zwaard, K. (2008). “A Holistic Approach for Establishing Content Authenticity and Maintaining Content Integrity in a Large OAIS Repository.” *Archiving 2008.*

PREMIS Editorial Committee. (2008). *PREMIS Data Dictionary for Preservation Metadata version 2.0.* <www.loc.gov/standards/premis/v2/premis-2-0.pdf>.

U.S. Government Printing Office (2005). *Authentication: Frequently Asked Questions.* <http://www.gpo.gov/authentication/faq/>.

U.S. Government Printing Office. (2005, October 13). *Authentication White Paper*. <http://www.gpo.gov/pdfs/authentication/authenticationwhitepaperfinal.pdf>.

U.S. Government Printing Office. (2005, May). *Future Digital System Concept of Operations*.
<http://www.gpo.gov/pdfs/fdsys-info/FDsys_ConOps_v2.0.pdf>.

van Diesen, R., van der Werf-Davelaar, IBM Netherlands. Koninklijke Bibliotheek. (2002). T. *Authenticity in a Digital Environment: Long Term Preservation Study Report Series Number 2.*

APPENDIX 1: How Digital Signatures Work

The signing software calculates a cryptographic hash value for the target content, and then encrypts that value with GPO's private key. Publicly-available software (i.e., plug-ins for Adobe Reader and Acrobat) uses GPO's public key to decrypt the hash value. It then recalculates the cryptographic hash value and compares that with what it unencrypted. If they match, the blue ribbon is displayed. Since no other entity has access to GPO's private key, a digital signature provides proof that the content was published by GPO. But how do users know that a given public key is really from GPO? GPO has proved its identity to a trusted third party certificate authority that has created a digital certificate. The software that checks the digital signature also verifies the digital certificate with the certificate authority, ensuring the public key cannot be spoofed.

This certification path ensures users that GPO was the source of the material and that GPO cannot later claim that it did not publish the material, which is call **non-repudiation**. Users must be connected to the Internet in order to have the ability to validate a digital signature on a PDF document. If a user is not connected to the Internet, the Certification Question Mark icon will display. Please see the *Authentication: Frequently Asked Questions* for more information. (U.S. Government Printing Office, 2005).

The chain of custody section above describes how GPO is recording all the significant events in the lifecycle of a package in the preservation metadata. The date and source of acquisition is also recorded. This creates a non-reputable chain of custody, which means that the originator and the publisher cannot later claim that it was not the source of the content. To facilitate the formalization of trust relationships between content creators and providers in the Federal Government, GPO implemented a Public Key Infrastructure (PKI). A PKI is the infrastructure necessary for the certification authority to link signer identities with their private keys.

APPENDIX 2: Events Captured in Metadata

Events triggered by Software

- Virus Check – inspection of file for malignant code
- Message Digest Calculation – calculation of cryptographic hash value
- Crypto Time Stamp Calculation – creation of a unique string calculated with cryptographic hash value and date of ingest for content files in the Archival Information Package (AIP)
- Ingestion – creation of an AIP
- ACP Creation – creation of an Access Content Package (ACP)
- Fixity Check – recalculation of cryptographic hash value and match against expected value in metadata
- Rendition Submitted – submission of an identified representation with the Submission Information Package (SIP)
- Rendition Creation – creation of new representation of content, usually derivatives for access, such as HTML.
- Digital Signature Assignment – addition of a digital signature on a PDF file
- Parsing – generation of descriptive metadata for search and creation of a new representation that has granules
- Deletion – removal of an object from the repository. A deletion event is always triggered by a “Rendition Deletion” or “Public Access Restriction” event.

Events triggered by authorized users

- Rendition Upload – addition of a new representation to an AIP or ACP, often as a preservation process
- Rendition Deletion – removal of a representation
- Submission – formal approval from the user for the repository to ingest and publish content
- Public Access Restriction – deletion of an ACP
- Replacement – replacement of another ACP with this ACP (for example, corrected content such as star prints in Congressional materials)
- Replaced – replacement of this ACP for another ACP
- Replacement Reversed – correction of an incorrect replacement of another ACP by this ACP.
- Unreplaced – correction of an incorrect replacement of this ACP by another ACP
- Delete Before Submission – deletion of an uploaded representation from the SIP before it was submitted for ingest.
- AIP Nominated for Deletion – nomination of an AIP for deletion
- AIP Approved for Deletion – approval of an AIP for deletion. Upon deletion of an AIP, the chain of custody information remains in the repository.