Preservation and Authentication of Government Information: Are We Ready for the 21st Century?

Judith C. Russell
U.S. Government Printing Office
Washington, DC, USA

Abstract

President James Madison once said: A popular government without popular information or the means of acquiring it, is but a prologue to a farce, or a tragedy, or perhaps both. It expresses well the primary reason for the creation of the Federal Depository Library Program (FDLP) and the essential mission of the Government Printing Office (GPO): Keeping America Informed.

There is a long tradition in the depository libraries and at GPO of providing public access to authentic government information and to preserving that information for future use. However, the rapid evolution of the Internet and the revolution it has brought to information access and dissemination is changing how we accomplish that mission now and will change it even more dramatically in the future.

GPO places an extremely high priority on assuring that the information we provide to the public directly and through the FDLP is authentic, and can be determined to be authentic, now and in the future. We are pursuing a combination of digital signatures and watermarks to ensure the authenticity of government information that is produced by the federal government and disseminated by GPO. We are also studying the techniques for data tagging, storage and migration that are most likely to ensure that future generations can access and use the content that we create and disseminate today.

GPO recognizes that providing permanent public access to electronic government publications is a significant commitment and one that will require us to replace the platforms and refresh the data periodically. The risk, of course, is that we don't know the characteristics of the hardware, operating systems and data structures that will be in use in five or ten years, let alone those that will be in use in fifty or one hundred years. GPO is meeting with technology leaders in industry and academia on a regular basis to look as far over the horizon as we can in order to make good decisions and anticipate future changes. We believe that the use of XML tagged data provides our best hope for migration of current information forward, but we continue to look at other options as well.

At the same time, GPO is actively engaged in initiatives to expand our partnership with the depository library community to preserve more than 2.2 million paper documents in the library collections and to digitize those publications for both preservation and improved public access.

Introduction

I would like to begin by giving you a little background on the Government Printing Office (GPO) and the Federal Depository Library Program (FDLP).

The chief executive officer (CEO) of the GPO is the Public Printer. Although GPO is a legislative branch agency, the Public Printer, like the Librarian of Congress, is nominated by the President and confirmed by the Senate. The current Public Printer is Bruce R. James. He came to GPO in December 2002. He appointed me to serve as Superintendent of Documents, and I returned to GPO in January 2003. (I also worked at GPO from 1991 to 1996.)

I hope that you are all familiar with the quotation from James Madison: A popular government without popular information or the means of acquiring it, is but a prologue to a farce, or a tragedy, or perhaps both. It expresses well the primary reason for the creation of the Federal Depository Library Program and the essential mission of GPO: Keeping America Informed.

The FDLP is actually older than the GPO itself. Bruce often marvels at the wisdom of the Founding Fathers, who in 1813 created the initial law requiring the deposit of federal government information throughout the country. To guard against the potential for a tyrannical central government, they insured that all citizens could exercise their rights to know about the actions of their government and at the same time benefit from the information compiled and created by their government. Their vision created a system that has lasted almost 200 years and has served us very well.

GPO opened its doors on the day of President Lincoln's first inaugural in March of 1861 as a printing facility for the Congress. Over time its roles expanded to provide printing for all of the government and eventually, after World War II, to procure printing from private sector printers on behalf of the government. In 1895, the Federal Depository Library Program was transferred to GPO from the Department of the Interior so the program could take advantage of access to publications as they were printed, and later procured, on behalf of Federal agencies. So GPO has three main lines of business: in-house printing, procuring printing, and information dissemination.

My title, in addition to the historical and statutory title of Superintendent of Documents, is Managing Director, Information Dissemination. I am responsible for all of GPO's information dissemination programs. I am the second librarian and the first woman to serve as Superintendent of Documents.

The largest of the information dissemination programs at GPO is the Federal Depository Library Program, which includes over 1250 libraries throughout the United States and its territories. Each library receives information from the GPO at no charge and in exchange provides no-fee public access to that information. This distributed system protects these valuable assets and ensures permanent public access. One of the strengths of the current system is that no single natural or man-made disaster can wipe out the collective record of our democracy because it is housed in hundreds of libraries throughout the nation.

The FDLP also funds *GPO Access*, an online service that provides free public access to a wide variety of government information in electronic form. Every day GPO adds current information such as the Congressional Record, the Federal Register, new bills introduced by the Congress, public laws signed by the President, and a variety of other government information. In an average day, the public downloads over 1.1 million government documents from GPO Access (http://www.gpoaccess.gov/), the equivalent of over 24 million typeset pages.

There is a long tradition in the depository libraries and at GPO of providing public access to authentic government information and to preserving that information for future use. However, the rapid evolution of the Internet and the revolution it has brought to information access and dissemination is changing how we accomplish that mission now and will change it even more dramatically in the future.

Authentication of Government Information

Bruce loves to do a dramatic story of a person ordering a document by sending a check to me (checks to purchase government documents from GPO are made payable to the Superintendent of Documents, although I never see them), then (according to Bruce) I go personally to pull the document off the shelf, place in it a Tyvek® envelope, address it and mail it. The individual receives the envelope and can't even rip it open, eventually cuts it open with scissors and once it is opened has the reasonable expectation that he has received an authentic government document. Bruce then asks how that same user will have the same degree of assurance that the document is authentic when it is an electronic publication.

Obviously, GPO Access is a trusted site, as are most agency websites, so someone who downloads the document from a trusted site has a reasonable expectation that she has an official copy. But what if she posts that document on her own website or e-mails it to a patron or colleague. How will that person know that he has received an authentic copy?

The solution that GPO is pursuing is a combination of digital signatures and watermarks. With a digital signature, any recipient of the electronic document can determine that the file is identical and unchanged since GPO certified that it was authentic by using a free software tool that permits the verification of the digital signature. GPO has almost completed its registration as a source for digital signatures and will begin signing new files on GPO Access in the near future. Over the coming year we expect to sign all of the Adobe Acrobat PDF files that we make available online.

A digital watermark provides a similar protection for paper copies printed from an electronic file and can even be embedded in the original printing of the document. The watermark is invisible to the naked eye, but can be enhanced by a relatively low cost machine (a few hundred dollars at present, but likely to become cheaper as this technique becomes more widely accepted). The watermark will be placed on each page of the documents and will then appear on fragments of a page and even on photocopies of the documents. GPO will probably ask some of our depository libraries to act as sites for the verification of watermarked government documents and provide the libraries with the equipment necessary to display the marks.

GPO feels that it is essential for electronic government documents to use these techniques for authentication so the public, the courts, and others can rely on the information now and in the future.

OMB/GPO Agreement for Executive Branch Printing Procurement

GPO is working on a number of projects to test various services that we may offer in the future. Perhaps the most exciting one, and certainly the one that has received the most press, is the June 2003 agreement between the Office of Management and Budget (OMB) and GPO.

The agreement, referred to as the OMB Compact, is truly a win for all concerned. It is an innovative approach to contracting for Executive Branch printing that is completely within GPO's statutory responsibility under Title 44 of the US Code. Under the agreement an agency may choose its own commercial printers using standard contracts issued by GPO. The publishing agency will pay GPO, and GPO will pay the printer, less a modest 3% fee to cover GPO's contracting and administrative costs. However, the printer will not be paid until my office has received two print copies and one electronic copy in the form of a print optimized Adobe Acrobat PDF file.

This will give GPO an electronic copy of each publication for preservation. The electronic file can be used to print future copies and a screen optimized PDF file can be derived from it for dissemination to the public, directly and through the FDLP. It will also give us two print copies to serve as "copies of last resort" so GPO can scan the paper copy to create a new digital copy in the future if the electronic copy is no longer useable due to changes in technology or other problems.

Partnership with the National Archives and Records Administration

Another part of GPO's efforts to create a reliable archive of government information is through a partnership with the National Archives and Records Administration (NARA). In August 2003, GPO and NARA signed an important Memorandum of Understanding (MOU). This contract makes GPO an official "archival affiliate" of NARA and all of the *GPO Access* databases are now considered the official archival copies of the publications they contain. The records are in the legal custody of NARA, but GPO will continue to maintain both the access and preservation copies.

Through this agreement, GPO and NARA, acting together, will ensure permanent public access to a wide range of important electronic government information. In addition, NARA continues to maintain GPO's historical collection of the publications distributed to depository libraries in paper and microfiche.

Preservation of Electronic Government Documents

Until *GPO Access* was launched in 1994, each depository library was responsible for preservation of the publications it received through the Federal Depository Library Program. Once GPO sealed a box and shipped it to the library, our job was done. The responsibility for both public access and preservation rested entirely with the library that received the materials.

GPO Access shifted the responsibility for public access and preservation of electronic government documents from the depository libraries back to the Government Printing Office. The agreement with NARA acknowledges the fact that GPO has assumed these responsibilities for the electronic depository library collection.

The electronic collection is already over 262,000 titles, with more than 157,000 available on *GPO Access* and another 105,000 available through links to agency and other websites. Wherever possible, GPO has entered into contracts with individual agencies to guarantee permanent public access to titles on the agency websites. Each of these contracts includes the provision that the materials will be transferred to GPO if the agency can no longer maintain access on its own website. Where no agreement exists, GPO harvests the electronic publications and places them in its own electronic archive. A persistent URL (PURL) is created, pointing to the document on the agency website. If the link fails in the future, the PURL will be redirected to the copy in the GPO electronic archive.

GPO recognizes that providing permanent public access to electronic government publications is a significant commitment and one that will require us to replace the platforms and refresh the data periodically. We are planning now for the next generation of *GPO Access* and expect to begin refreshing the current databases this year in preparation for migration to the new platform. Although the recent material on *GPO Access* is already tagged using XML, the older databases are not, so we will code those older databases to prepare them for the new platform.

The risk, of course, is that we don't know the characteristics of the hardware, operating systems and data structures that will be in use in five or ten years, let alone those that will be in use in fifty or one hundred years. All of us know examples of data that was produced ten years ago that is embedded in proprietary file formats, dependent on operating systems that are now obsolete, or stored on media that can only accessed on hardware that is so outdated that it is rare to find a piece of equipment outside a museum.

We have depository libraries that are keeping old computers running Windows 3.1 in order to access data on CD-ROMs published in the early 1990s. GPO is working with the library community now to assess the data that is at risk and develop a plan to migrate it out of proprietary file formats and free it from software that is dependent on obsolete platforms. Where that is not possible, we will work with the publishing agency to reissue the data is a form that is more accessible for current use and more sustainable for future access. This is one example of the new responsibilities that GPO has assumed and also a warning to all of us that the decisions we make today, in good faith, based on the best information currently available, are likely to have unanticipated consequences in the future.

GPO is meeting with technology leaders in industry and academia on a regular basis to look as far over the horizon as we can in order to make good decisions and anticipate future changes. We believe that the use of XML tagged data provides our best hope for migration of current information forward, but we continue to look at other options as well. Through the National Information Standards Organization (NISO) we are participating in development of standards that will help us manage electronic files for permanent public access, such as PDF-A, the archival version of the Adobe Acrobat portable document format.

And, as noted in the description of the OMB Compact, we are also preserving paper copies of many of these documents, just in case.

Preservation of Legacy Collections of Government Documents

GPO is working on preservation of the tangible government documents collection as well as the electronic collection. During the past year significant progress has been made on the issue of managing our legacy print and microfiche collections. There are three related initiatives that are underway simultaneously.

Shared Repositories

First is a movement toward shared repositories, or shared housing agreements, that would allow two or more libraries to eliminate some of the redundancy between or among their collections. These initiatives are still in the early stages, but they are very important since they will help us move toward a smaller number of comprehensive sets that can be more readily preserved.

Each of our 1280 depository libraries is not going to preserve its entire collection of tangible government documents. Even the 53 regional depository libraries, which have the most comprehensive collections, cannot afford to actively preserve all of their government documents. But we do need to decide as a community how many sets of tangible Federal documents should be preserved and take the necessary steps to establish consolidated, comprehensive collections, so we can actively preserve the materials. The Center for Research Libraries (CRL) is developing a report on the characteristics of trusted repositories that will help us evaluate the options and opportunities.

Collection of Last Resort

Second is the decision by GPO to establish a collection of last resort. At a minimum this will become, over time, a comprehensive collection of tangible and electronic titles that will serve as a dark archive to backstop the collections in the regional libraries or the repositories as they are established.

Based on advice from the library community, GPO has developed a preliminary plan for the collection of last resort, which is currently being reviewed. The plan includes in the collection of last resort the two print copies of each document received as part of the OMB Compact. It also anticipates the GPO will create and preserve two print copies for other born digital publications, in case the electronic copy cannot be used in the future due to changes in technology or other problems.

Digitization of the Legacy Collection

Third, and essential to the other two, is the decision by the Association of Research Libraries (ARL) to collaborate with GPO, and ultimately with the entire library community, on a national digitization plan, so that we can coordinate our efforts to digitize a complete legacy collection of U.S. government documents and make sure that the documents are available, in the public domain, for permanent public access. This is no small task. We estimate that there are 2.2 million paper titles, representing over 60 million pages to be converted, not counting the microfiche collections.

The ARL proposal is one of several related efforts that, together, will make it possible to accomplish this goal within the next few years. There is one regional depository library that is willing to allow its collection to be used for a comprehensive digitization initiative and willing to provide space for scanners and personnel in its facility. The National Agricultural Library is interested in working with GPO and the land grant universities to digitize the entire legacy collection of USDA publications. We are having discussions with Congress and other agencies about their desire to have accessible online collections of their legacy publications.

There are many initiatives that will contribute to this effort. It will be a collaborative effort. As I see it, a number of libraries will actively digitize materials, based on established priorities or local needs, while other libraries will collaborate to support the digitization specific materials. A variety of government and foundation grants and private sector partners will facilitate this effort.

GPO's roles will be to coordinate the effort, assist in the establishment and coordination of standards, serve as a trusted repository for preservation and access (in addition to any other places that the materials might be held), certify and authenticate the electronic files, and ensure that there is appropriate cataloging and metadata for the items in the collection.

GPO has requested funding in FY 2005 to perform OCR on the digitized files and output XML tagged data that can be used for preservation, public access and for print-on-demand. Thus, whatever OCR scanning is done by individual libraries or other partners, we can ensure that the preservation and access collection maintained by GPO is consistently tagged, making it a true collection, not just a random assortment of electronic files. We will also continue to work with Congress, the Judiciary and Federal agencies to get them to participate and, where appropriate, to certify the files as official copies.

This is an extremely shorthand description of a complex set of actions which together will help us preserve a reasonable number of copies of the tangible artifacts as well as to create and maintain a comprehensive, digital, public domain collection for preservation and access. The availability of these tangible and electronic collections will allow all depository libraries to manage their collections more effectively, substituting electronic copies for tangible copies if they wish to do so. And it will ensure that the legacy collections now available only in print and microform are fully a part of the electronic library collection of the future.

Conclusion

GPO places an extremely high priority on assuring that the information we provide to the public directly and through the Federal Depository Library Program is authentic, and can be determined to be authentic, now and in the future. We recognize the need to expand our partnership with the depository library community to preserve the paper and other tangible documents in the library collections and to digitize those publications for both preservation and improved public access.

The issues of authentication and preservation are relevant to future content, whether derived from print or born digital publications, as well as to the management of the legacy collections in their original and increasingly digitized forms. We must manage our legacy paper and microfiche document collections and decide how many tangible copies we should (and can) actively preserve. We must improve access and ensure preservation of the content of the legacy collections through digitization and preparation of the appropriate cataloging records and other metadata. We also need to be printing out and preserving two or more paper copies of born digital publications in case the current electronic formats cannot be sustained.

Bruce frequently says that he did not come to Washington to run a printing plant. He came to address the challenges of permanent public access to government information. He sees GPO's primary mission as information management and dissemination, with printing as one way to accomplish that mission, but by no means the only way.

GPO recognizes that authenticating and providing permanent public access to electronic and tangible government publications is a significant commitment and a substantial challenge. We are proud to share that responsibility with our partners in the library community. We are ready for the challenges and welcome the opportunity to continue to Keep America Informed through the 21st Century and beyond.

Biography

Judith C. Russell is the Managing Director, Information Dissemination and Superintendent of Documents at the U.S. Government Printing Office (GPO).

Russell previously served as deputy director of the U.S. National Commission on Libraries and Information Science (NCLIS) and as director of the Office of Electronic Information Dissemination Services and Federal Depository Library Program at GPO. She holds a Master of Science in Library Science (MSLS) from The Catholic University of America in Washington, DC.