Council Session: GPO Digital Content Forum –

CD-ROM Data Migration

U.S. GOVERNMENT PRINTING OFFICE
KEEPING AMERICA INFORMED

Monday, April 3, 2006 3:30 – 5:00

Background of Presentation:
The presentation outlined below discussed CD-ROM Data Migration, as presented by Lisa Russell, Content Manager, Planning and Development Content Management. This served as one portion of the Council Session entitled, "GPO Digital Content Forum." Facilitated by Richard Davis, Director, Library Services and Content Management, this session offered four separate presentations on:

- **Web Harvesting**
- **CD-ROM Migration**
- **Digitized Content and**
- **Digitized Content Specifications**

CD-ROM Data Migration

Key Discussion Points - GPO is developing a strategy to maintain access to the content of depository CD-ROMs over time to ensure permanent public access to discs that utilize older technology.

- Depository libraries currently attempt to maintain older computers in order to access government information content on older formats that may not run on newer technologies.
- GPO is addressing this problem by evaluating options for preserving access to the content of older CD-ROMs. Other media are also issues, but GPO is starting with CD-ROMs to start to devise a strategy and will address the other media at a later date.
- Addressing at-risk content that may be lost is the first priority.
- Data preservation may include refreshing data, data migration, and/or emulation strategies.

Summary:

- As technology changes, it becomes more difficult to access content on older media.
  - The hardware and software used for these products become obsolete, and it becomes difficult to maintain older computers needed to access these products.

  - Proprietary software requires maintenance of software licenses or may not be supported by vendors as they develop newer versions of the software, making it difficult to access content on older products. Some software vendors have gone out of business, leaving GPO and depository libraries without technical support for their products.
- GPO is taking steps to address the problem by identifying issues, identifying potential solutions, evaluating potential solutions, developing a strategy, and implementing a solution.
- Potential solutions include refreshing data, data migration, and emulation technology.
  - Refreshing data is moving data to new storage media periodically.
  - Data migration is the periodic transformation of files into a new digital format, ensuring compatibility between file formats and applications.
  - Emulation allows obsolete systems to be run on future systems, maintaining the use and application of the original documents with the original software.
  - A final strategy may include some or all of these solutions to ensure access to content.
- GPO staff conducted a review of a sample of CD-ROMs included in an inventory spreadsheet provided by the Regional depository at the University of Kentucky.
  - Three agencies were used for the sample: Department of Education, Department of Justice, and United States Geological Survey (USGS).
  - Discs were prioritized based on whether the content was found on the Internet in order to ensure that content that was not available elsewhere was addressed first.
  - Lessons learned from the sample will be used to develop a strategy for handling discs from other agencies.
- GPO is working with other stakeholders to address data migration issues.
  - GPO is an active participant in the Government Information Preservation Working Group (GIPWoG). (See http://www.itl.nist.gov/div895/gipwog/index.html for more information on GIPWoG.)
  - GPO is working with the National Institute on Standards and Technology (NIST) on a pilot project to test emulation technology as a potential solution to make older discs accessible.
- There was a brief demonstration of the emulation pilot that is GPO is conducting with NIST.

U.S. GOVERNMENT
PRINTING OFFICE
KEEPING AMERICA INFORMED

Digital Content Forum: CD-ROM Migration - Q&A

Monday, April 3, 2006 3:30 – 5:00

Question 1

Who decides how harvested content is added to the FDLP? Is there the subjective human element that will examine these documents, or is there some criteria that is triggered by this project once they've been ingested?

Summarized Responses

- The subjective human element is involved by defining the roles and parameters that are associated with the crawling process. Part of the reason we're conducting multiple crawls is that we are continuing to have human beings involved in that refinement process to try to improve the results each time.
- We are receiving questions from the two companies about what is in scope and what is out of scope. Experts at GPO are making those decisions, and the companies are making rules based on those decisions. We hope the final product will be a good set of rules to make those types of decisions.

Question 2

How is the crawler going to distinguish changes over a period of time as content is edited?

Summarized Responses

- The expectation is that it will keep a record of what it has previously harvested when it repeats a crawl. If it finds a duplicate file name, it will examine the document to see if there is a change in either the file date or the number of bits and bytes. If either of those has changed, the crawler will consider it a new version.
- We expect some information from the two companies about the effect of what they are doing in terms of identifying new versions. For example, an agency may have the current issue of Journal X available online and use the same file name when a new current issue is posted, changing the file name of the previous issue to something else. The crawler should be able to detect that the file name is the same and that the file is in the same folder, but the date and number of bits and bytes have changed, and therefore retrieve it as a different object.
- The history of the crawl will be tracked.

Question 3

What is the quality control process on the OCRs for the PDF files?

Summarized Responses

- The process of the OCR component of creating the PDF is a function of what we are calling Release 0. We have a white paper authored by one of GPO's technical experts that will be posted online. In our testing we have found that a major factor in the accuracy was the resolutions of scanning that was used – 300 DPI versus 400 DPI versus 600 DPI. Based on the study we conducted, we are able to get better than 99 percent accuracy.

Comment

The CD-ROM project deals with a very vexing problem. I especially like the idea of posting a list of CD-ROMs for the public to see, because there may be things going on in individual libraries with some of the content that may relate to the rescue of the most seriously at-risk critical content. I think it is an impressive project.

Summarized Responses

- Just a follow up, some of these discs, particularly ones from the early 1990s, have content that is intrinsically tied to proprietary software that is no longer supported. It will be fantastic if we can make that content accessible into the future.

Question 4

Has any approach been made to libraries that have posted digitization projects to the registry? That would be a good group of people to solicit or start a conversation about what the specifications are and the possibility of incorporating or harvesting their digitized files into the system.

Summarized Responses

- We have not yet made contact, but that is one of the ways we hope to use the registry.

Question 5

Regarding the CD-ROM project, we had a demonstration of emulation. Are you looking at migrating or refreshing data as well? Are those on the table?

Summarized Responses

- Data refreshing and migrating are also on the table. The outcome of the project may be a strategy that is some combination of all three. For example, maybe for some file formats data migration may be a straightforward method of ensuring access, whereas for something that utilizes a software package emulation may be more effective. We hope to clarify some of those issues as we work through the process.

Question 6

Are there plans for authentication of the digitized content?

Summarized Responses

- Yes, GPO intends to authenticate all of the PDF files that we have on *GPO Access* or in the electronic archive, including those that we create ourselves through digitization.

Question 7

VHS is being phased out, and videos are being put on DVD and CD instead. Is GPO addressing this?

Summarized Responses

- The National Digital Information Infrastructure project at the Library of Congress is looking at audio and video. GPO has really stressed that we are starting with print publication, but obviously some of the CD-ROMs will take us into other media. We know that at some point we will need to address things that were distributed on VHS or other media. The plan for the Future Digital System is to include audio and video materials.
- We are working on the print first, and then we will be looking at different specifications as we address other media. That is one of the areas that we are discussing with Library of Congress and NARA, because both of them are also dealing with other media.

Question 8

What standard has been used for digitization? What kind of OCR software has been used for the PDFs?

Summarized Responses

- It is too early to say what the final solution will be in the Release 0 processing because we just received approval from JCP. Now we have the ability to formulate or validate work processes and find out what works best, based on how content will ultimate be searched, posted, etc. There may be one solution, or there may be multiple solutions. It is too early to say there is one solution and this is what it is, but that evaluation will be part of the pilot.

Question 9

One of the CD-ROM titles you demonstrated and some of the others on the screen were not from the three agencies identified for the sample. How were those chosen?

Summarized Responses

- Historically, GPO has not had a collection, so we generally do not have copies of the CD-ROMs that have been distributed to depositories. When we are ready to implement a strategy, we may be requesting copies of CD-ROMs from some of you in order to have discs for the project. Fortunately, around the time we began working with NIST on the emulation pilot, we found a small collection of older CD-ROMs at GPO. We selected our test discs from that collection, rather than trying to acquire copies of titles from our sample of Education, Justice, and USGS discs.

Comment

I have a follow-up on the scanner. We are doing some scanning, and our OCRs are at about 98 percent, which is not good enough for a couple of the things we are scanning. As this develops, it would be great if GPO could share information with the community. Posting information on *GPO Access* about the current configuration GPO is using could be a leading indicator, so to speak, for the rest of us.

Summarized Responses

- The white paper that we have distributed internally includes some of the results we have gotten in our scanning. We are not at the point of saying that there is a preferred or selected scanner, but we have gathered a good deal of information over the past 12 months while researching not only the work process, but the specifications themselves.

Question 10

Could somebody describe briefly the relationship between the TIFF and the PDF? You scan it as a TIFF and then it becomes a PDF? How does that work?

Summarized Responses

We start with these TIFF images – a single TIFF image for every page of the document, including all the blanks. A unique ID is established at the document level, and there are sequential numbers for each TIFF image that correspond from the front of the book to the back of the book. For a 100-page document, there will be 100 TIFF images.

When the OCR software processes the TIFF images, it reads them and enables the capability to create PDF files for each individual page or a cumulative PDF file that has all 100 pages, for example. The OCR text can be saved behind the scenes or up front where it can actually be seen. That's a real brief explanation of the process. We start with the scanned TIFFs, and then create that searchable file using the OCR software.

Comment

To add to that, the key for preservation is that those TIFF files then become the master files that are considered the preservation copy. The PDFs are the derivatives that are the access files. The key is that those TIFF files are richer files, which is why they're kept as the master copy.

Summarized Responses

- That is correct. The creation of the Submission Information Package is really the beginning or the entry point into the Future Digital System. The TIFF images and the accompanying metadata are what is preserved. Then as technologies change, we can repurpose those TIFF images into whatever format is needed. The files are huge, so they require plenty of storage space.

Question 11

Will the TIFF files be stored on servers elsewhere than GPO for preservation purposes, so they will be somewhere else if something happens at GPO?

Another comment on OCR, one problem that we have had is that as we go farther back into our legacy documents into the varied historic documents from the 1700s and 1800, the OCR becomes less effective.

Summarized Responses

- The files will be stored on servers outside GPO for preservation purposes. The final hierarchical storage process solution will be through the Future Digital System, and the core requirements of that solution are part of the RFP document that was recently released. The final pieces will be developed and implemented as part of the overall architecture that will be developed when the master integrator is selected. In the interim, some of the requirements documents for storage will be good resources.
- Regarding the OCR, you are correct. In GPO's workflow, every document is looked at individually. In some of the older material – for example, newsprint that is faded, discolored, wrinkled, and crinkled – there may be instances where re-keying is the solution to achieve the necessary accuracy. Then we get into issues of official authentic content, the goals of unstructured search to find the document, etc.