Council Session: GPO Digital Content Forum –
Digitized Content Specifications

U.S. GOVERNMENT PRINTING OFFICE
KEEPING AMERICA INFORMED

Monday, April 3, 2006 3:30 – 5:00

Background of Presentation:

GPO will provide the following for Digitized "Converted" Content:

- Consultation and Collaboration
- Retrospective document scanning
- Metadata creation
- Access
- Preservation services

Core digitization documentation:

- FDsys Operational Specification for Converted Content (Version 3.3)
- Specifications and Metrics for Quality Control of Converted Content (Version 1.1)

Summary:

- GPO staff and external service providers "including contractors, library partners, and federal agencies" will provide converted content to the Future Digital System. The end product of conversion is a Submission Information Package (SIP). The SIP must be produced at a level of quality that is adequate to support preservation as well as future iterations of derivative products.

- GPO's digitization specifications, standards, and workflow will continue to evolve and improve as technological advancements occur.

Council Session: GPO Digital Content Forum –
Digitized Content Specifications

Monday, April 3, 2006 3:30-5:00

Background of Presentation:

The presentation outlined below discussed Digitized Content Specifications, as presented by Ted Priebe, Director, Library Planning and Development. 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

Digitized Content

Key Discussion Points - GPO is digitizing fugitive publications for dissemination to the Federal Depository Libraries. These publications are available for access through the Catalog of U.S. Government Publications.

- Publications for this presentation are monographs
- Scanning was done to FDsys Specifications for Converted Content
- Cataloging has been done for these publications and is available through GPO's new ILS and OCLC
- The publications have been converted to PDF files—to meet public access needs
- The files shown were originally digitized and saved as TIFF images to meet preservation needs for the future and access derivative files were produced from the TIFF files

Summary:

- GPO continues to work on developing digitization specifications and requirements to fulfill current access and future preservation needs for published Federal information products. A pilot project will be starting to digitize not only the fugitive publications currently processed, but also to proceed with systematic digitization of all published Federal information products in the Federal Depository Library Program collections.

- The following published Federal information products are produced in a fully searchable PDF format.
- The Federal information products reviewed for this presentation were:

  Oil and Gas Management Plan: Big Thicket National Preserve, 2006

  Drinking Water Inspector's Field Reference, 2003 Edition

  H.R. 10499: A bill to amend the Social Security Act to liberalize benefits under the old age, survivors, and disability insurance program….

  Advice on Providing Additional GSP Benefits for Least Developed Countries

  State of the Watershed: Water Quality of Boulder Creek, Colorado

Council Session: GPO Digital Content Forum –
Digitized Content Specifications

U.S. GOVERNMENT PRINTING OFFICE
KEEPING AMERICA INFORMED

Monday, April 3, 2006 3:30-5:00

Background
Ted Priebe was one of the speakers for the Council Session: GPO Digital Content Forum on April 3, 2006.

Question 1:
Clarification on the resolution of the PDF files that were shown: are they screen-optimized PDF files? If so, are they at a much lower resolution or down sampling than the actual live images?

Response:
Yes the screen optimized PDF files are at a much lower resolution that the preservation masters "TIFF's". GPO can produce press optimized PDF files at a much higher resolution for print quality.

Question 2:
Question regarding quality control for the OCR processing of the PDF files.

Response:
We have a white paper that was authored by one of GPO's technical experts, and it will be posted on the FDsys projects portion of GPO's web site. What we've found in our testing is that we were able to achieve greater than 99 percent accuracy. One of the biggest reasons for that of that is the resolution of scanning; we compared 300 DPI versus 400 DPI versus 600 DPI.

Question 3:
Could you describe the relationship between the TIFF images and the PDF, based on the process that when you scan it and it's a TIFF, but then it becomes a PDF?

Response:
GPO's specifications call for digitizing single TIFF images for every page of the document "including all the blanks". There's a unique ID that is established at the document level, and then there are sequential numbers for each TIFF image that correspond from the front of the book to the back of the book. So if it's a 100-page document, you have 100 TIFF images. When those TIFF images are processed through OCR software, it reads them and enables the capability for you to create PDF files for each individual page or a cumulative PDF file that has all 100 pages. We currently save that OCR text behind the scanned image for unstructured search capability.

Question 4:
Are the TIFF files the master files, which would be considered the preservation master? Would the PDFs "derivatives" be considered the Access files?

Response:
The TIFF images with corresponding metadata “brief bibliographic information” are considered the preservation masters (these make up the submission information package) that will flow into FDsys and be preserved.

The PDF files that are derivatives of the TIFF images are considered the Access files.

Question 5:

Regarding the TIFF images, will GPO have these images stored with redundancy in case something where to happen to the central storage location?

Response:
Yes, FDsys has requirements for hierarchal storage management “redundancy” processes within its Requirements Document (RD 2.0) that will be incorporated by the Master Integrator.