# 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.