U.S. GOVERNMENT PRINTING OFFICE
KEEPING AMERICA INFORMED

Digitization of the FDLP Legacy Collection

Contents:

I. BRIEFING TOPIC: Digitization of the FDLP Legacy Collection

I.1. SETTING THE STAGE

GPO is working with the library community on a national digitization plan with the goal of digitizing a complete legacy collection of tangible U.S. Government publications held in libraries participating in the Federal Depository Library Program (FDLP). The objective is to ensure that the digital collection is available, in the public domain, for no-fee permanent public access through the FDLP. The project will ensure that the collection is digitally reformatted for preservation purposes. The digital preservation masters and the associated metadata will be preserved in the GPO electronic archive (in addition to any other places that the materials might be held) and there will be no-fee public access to the content through derivative files on *GPO Access*.

The materials in the legacy collection include approximately 2.2 million print documents, over 60 million pages, of mostly textual material, but with some additional information, such as charts, diagrams, and photographs as included in Federal publications since the beginning of the Federal Depository Library Program, dating back essentially to 1813. Materials not included in this initial project are the maps, microfiche, and audio/visual material that have been distributed through the FDLP. These will be digitized at a later date. The project will include digitization of depository resources by libraries participating in the FDLP. Not all institutions will be able to participate, and not all participating institutions will have the same economic resources available for the digitization project. At this point a comprehensive plan for the digitization of the legacy collection does not exist, although a number of steps have been taken to assist in developing the plan.

The following steps have been taken to assist in developing the plan:

- The first in a series of meetings with experts on digital preservation was held in GPO in March 2004. The *Report of the Meeting of Experts on Digital Preservation: Digital Preservation Masters*, which contains scanning specifications, is located on GPO Access at http://www.gpoaccess.gov/about/reports/preservation.html. A chart that summarizes scanning specifications follows in this section of the briefing book.
- The second meeting of experts on digital preservation was held in June 2004 at GPO. The summary report, *Meeting of the Experts on Digital Preservation: Metadata Specifications* has been prepared, but not yet shared with digital experts participating in the in the meeting.
- A Digitization Priority Survey has been held to identify the most important titles to be digitized first in the project. Depository and non-depository libraries were invited to participate in the survey. While this was a non-binding survey, it should help GPO and individual libraries make decisions about titles for digitization. These results allow GPO and its potential digitization partners to identify overall priorities, focusing attention on high visibility titles and providing a resource for consultation when institutions are planning digitization projects. Information about the survey is located at http://www.gpoaccess.gov/legacy/priorities/index.html. The survey is now closed and a summary of the results is included in this section of the briefing book.
- Initial discussions have begun with ALA GODORT for a partnership with the Government Information Technology Committee (GITCO) on a digital registry partnership. This registry will include information from a number of different registries as we are hoping to also incorporate the required elements from the OCLC/DLF Registry, ARL registry, and IMLS.

Information about library digital scanning projects and best practices should be shared with GPO by Council to ensure the digital plan we are developing is clear and beneficial to all. Specifically, GPO needs feedback from council on a list of essential elements for a metadata scheme, identification of essential

elements for authenticity, officialness, and certification of resources, feedback on the necessary parameters for versioning, review of the scanning specifications for application to the differing libraries participating in the FDLP, review of recommendations for the first documents to digitize, and suggestions of accurate and efficient approaches to clearly identify the scope and breadth of the materials to be digitized for the project.

I.2. NEW INFORMATION

New information that becomes available on this topic between the time of this writing and the presentation at the conference will be presented at that time.

I.3. MICRO RECAP

Digitization of the historical legacy collection, materials held in the library collections located in the depository libraries throughout the United States is the focus of the discussion of this fact sheet. The challenges associated with this project, to develop a plan to digitize over 2.2 million documents, 60 million print pages, is tremendous. GPO will be working cooperatively with depository libraries and others interested in ensuring the content of the collection remains in the public domain. While GPO may put forth a plan, it will only be effective with participation and an exchange of information with the libraries in the depository community, and in also with the larger information community. There are a number of issues associated with this project, including size of the collection, authenticity and official status of publications to be digitized, scanning, versioning and metadata requirements, and location of the digital files.

As identification of material prior to 1976 is needed to determine the actual size of the project (a better estimate than 2.2 million documents), it is essential we have an accurate and time effective way to do this. While GPO maintains ownership over this material even though it is located in depository library collections, actually acquiring the material will prove difficult, especially for the older and more fragile resources.

Information about the digital objects is needed to ensure present and future access. There are many different metadata schemas available, but it is essential GPO choose a schema that will allow maximum flexibility for partner libraries to do the metadata, and yet provide sufficient detail to allow the digital files to be used for many purposes, including providing current access, repurposing files for print-on-demand products, ensuring permanent public access.

Minimum scanning specifications have been developed in coordination with a group of digital experts. These specifications are focused on print documents that have been distributed through the program. When possible, a recommended and minimum specification was identified to allow maximum flexibility for the institutions digitizing the resources. Not all institutions will be able to participate, and not all digitization projects will meet the scanning requirements. Some material will only be scanned for access, thus some materials will not have preservation level files for permanent public access.

II. REVISED ASSUMPTIONS:

The digitization project of the Historical Legacy Collection will develop a complete collection of tangible U.S. Government publications held in libraries participating in the FDLP.

II.1. The digitized resources resulting from this effort will be available in the public domain for no-fee permanent public access.

II.2. The primary result of the process will be digital preservation master quality files from which access copies and other derivatives will be made.

II.3. The resources for the first phase of the project include the approximately 2.2 million print publications, over 60 million pages of mostly textual materials, but with some additional pieces, including charts, diagrams, and photographs that are included in the publications.

II.4. The second phase of the project will include the maps, microfiche, audio/visual material, and tangible electronic publications distributed to the libraries under the FDLP.

II.5. Not all depository libraries will be able to contribute to the digitization effort, but will share in the result.

II.6. GPO's main role is facilitating the digitization effort and to provide specifications to use in developing digital projects, based on best practice.

II.7. The digital preservation master files produced under this project will be part of the National Collection of U.S. Government Publications, and access derivatives will be available through the FDLP Electronic Collection.

II.8. Depository libraries, non-depository institutions, and Federal agencies will contribute to the digitization project by:

a. digitizing their own publications

b. acquiring copies of necessary resources (such as missing copies) and digitizing the material

c. providing additional information such as metadata elements or additional cataloging for already digitized products

d. providing a server or hosting location for the digitized material

II.9. GPO will obtain digital preservation master files from the participating libraries for placement in the National Collection. The digitizing institution may also make the files available from their Web sites or servers.

II.10. GPO will use the preservation master files to ensure permanent public access, and develop derivative files, such as PDF and ASCII, to provide for access and Print-On-Demand.

II.11. There will be different levels of authentication for the publications digitized, depending in part on what the material is and how it was obtained:

a. Distributed through the FDLP and maintained in the library's collection

b. Material obtained from a Federal agency, but considered by GPO to be a fugitive document (never distributed through the FDLP)

c. Material obtained from a commercial vendor, or private party; did not come through the FDLP or a Federal agency

II.12. Libraries will be allowed to substitute access to the copies in the FDLP Electronic Collection for tangible copies housed in their library collections, thus freeing up space and reducing staff time necessary to maintain the collection.

II.13. Each of the digitization efforts will need to provide sufficient metadata to allow the publication to be managed and made accessible. The lowest level of metadata required has not yet been determined.

III. QUESTIONS TO COUNCIL, WITH COUNCIL DISCUSSION

III.1 QUESTION: What approach should GPO take to the titles prioritized in the recent survey:

a. digitize a limited set of the top 20 priority items covering a 3 to 5 year period, or
b. digitize a smaller number of priority titles with comprehensive coverage?

DISCUSSION BY COUNCIL
[tabled and not revisited]

III.2 QUESTION: Can this be effectively accomplished in a standardized way by many participating institutions?

a. Is there a core group that GPO should do internally?

DISCUSSION BY COUNCIL
GPO should do a list of 25 internally.

III.3 QUESTION: What should happen to the digital preservation master produced by the digitization effort:

a. distributed model with preservation masters held at various institutions
b. a centralized model with all preservation masters held by GPO, or
c. a comprehensive set of masters at GPO with some copies at other institutions?

DISCUSSION BY COUNCIL
The C scenario reflects the dark archive as well as several light archives, assuring redundancy.

IV. QUESTIONS FROM COUNCIL ADDRESSED AT THE MEETING

IV.1 QUESTION: Is C assuming that there is one set of master copies and there are selective copies elsewhere? Should there be a D option in which there is a set of master copies somewhere and then other copies elsewhere as well?

RESPONSE: If this is done as distributed digitization, certainly the digitizing institution will have the right to keep a digital preservation master file. It is important that there be some kind of government responsibility for a set of preservation masters. Our sense is that there needs to be a dark archive.

IV.2 QUESTION: Would a master set exist somewhere, with an additional copy of everything somewhere else?

RESPONSE: We're already assuming that we would have two dark collections. We would have a redundant backup, because you don't ever keep electronic files without some redundancy. But there may also be fragments of the collection in individual institutions that want to have some of those master files.

IV.3 QUESTION: Are we talking about a central coordinating authority providing the most completely cost-effective dissemination of locator services or are we talking about two different entities doing this?

RESPONSE: GPO sees its responsibility as insuring that there is permanent public access to the derivative files and preservation of the master file. There are different ways of doing that: with NARA, with publishing agencies, with partners in depository libraries. We need to spread the risk among the different institutions, and provide redundancy at the same time.

IV.4 QUESTION: Does the FDLP electronic collection form a part of this legacy collection?

RESPONSE: Yes.

IV.5 QUESTION: Following on the preservation aspect of Mike Wash's presentation yesterday, does this include refreshing the files in some way?

RESPONSE: That is definitely essential to preservation. The assumption is that the files will be migrated forward over time.

IV.6 QUESTION: Some of the constraints we're talking are based on the cost of storage, and cost of storage is getting cheaper. I would like to see very wide distribution to partnering libraries, or to institutions that in a way mimic the FDLP. If we have lots of copies, not 2 and not 4, then we can assuage some of the fears that materials will be removed for political reasons.

RESPONSE: We are assuming that there will be lots of access copies.

V. AUDIENCE QUESTIONS ADDRESSED AT THE MEETING

The facilitator of the Council sessions accepted questions from the audience written on GPO-supplied cards. Fifteen of sixteen questions were answered during the Council session. Those questions and their answers are summarized below. One question held to answer at a later date, either because of time constraints or the need for a subject matter specialist to provide a more detailed answer, follows the questions answered during the session.

V.1 **QUESTION:** Would someone please explain the business plan for digitizing the legacy collection? I don't understand how it will be financed initially and in perpetuity.

RESPONSE: There are a variety of means of funding, some using existing appropriations, and some involve getting permission to expend money from prior year appropriations. We've had some recent requests from the Congress to have specific materials digitized for their own use and added to GPO Access. And that could potentially be funded out of the Congressional printing and binding appropriation and not from the Depository appropriation.

We may be able to use some of the retained earnings in our revolving fund to invest in this project. And if GPO moves out of the existing building into more modern, much smaller facilities, and if that land is retained as federal land but developed in some way, then that could provide an income stream independent of the appropriations process.

V.2 **QUESTION:** In digitizing the legacy question will priority be given to documents most at risk either because of comparative rarity or because of preservation concerns?

RESPONSE: We haven't yet determined the priorities. But that is certainly one of the criteria that have been put on the table.

V.3 **QUESTION:** I would like to hear more about the project to migrate data from CDs.

RESPONSE: We are working with GITCO and LITA at ALA. We've been doing an assessment of the state of the current CDs. At some point a report will come back to GPO which says here are CDs where we can extract the data and migrate it out and put it on another platform, much as we did with floppy disks. There may be others where the data is deeply embedded in proprietary formats and we'll have to go back to the publishing agency and see if they can help us. They may have already produced that data in another format, and we might be able to substitute or replace that data. But we're in the analysis process now.

V.4 **QUESTION:** Priorities for digitization should be based on least accessible, and more endangered material, not solely on format. Microfiche and audio/visual and CD-ROM material should not be relegated to second tier.

RESPONSE: The challenges of digitizing the material that's available on microfiche are significant, because often we will need to find a print original for each document. We have talked to NARA about the extent to which we can use the national archives

collections to find the originals. We may also need to go back to the publishing agencies for copies.

We're not tackling microfiche first because there's a significant effort involved in just finding the materials, compared with print materials.

The Department of Energy is undertaking a project to get information locked up in microfiche out in full text searchable format. There is evolving technology which allows microfiche to be digitized at a very low cost. It's not the same quality you get from an original paper copy. But it is sufficient that it is readable and it is OCR and therefore searchable.

V.5 **QUESTION:** GPO does not want to compete with commercial publishers and also wants to sell enhanced data. How will these conditions affect the quality of the product? For example, the serial set has already been digitized by two commercial firms. To what extent will GPO's or partners' digitization of a serial set compete with commercial editions in terms of quality of images, indexing, metadata, advance searching capabilities, et cetera?

RESPONSE: It's difficult to say because we aren't at that point yet. We have said consistently that we expect to create a comprehensive collection. The serial set will be part of that collection and it will receive treatment comparable to the other items in the collection. Private sector indexing might be far superior to what we are proposing, because we aren't looking at the kind of editorial indexing that both firms are doing. We have spoken with both firms very openly about our program. We asked both of them if they would consider ways to try to partner with us to put the parts of the collection that they digitize in the public domain so that we don't redigitize a third copy. Neither of them has said they wouldn't but neither of them have said they would. It's difficult obviously when there are two teams out there that are competing.

Many of the products that have been developed in the private sector have been developed because the government was not doing a good job of managing its own information. If we get our act together, then that's what the government should be doing. That leaves lots of room for innovation by the private sector for enhanced indexing and for a combination of this government material with material from copyrighted and other resources. It opens a vast array of material for them to develop. But we should not redline material and say we're not going to have this in our comprehensive collection merely because a private sector company is also exploiting that material.

V.6 **QUESTION:** How will you handle dynamic databases where documents are created on the floor?

RESPONSE: We'd love to know the answer to that question. That's why we're starting this whole versioning discussion on Wednesday morning. We recognize that different kinds of data are going to have to be captured at different periods in order to have a representation of what the data was at that point in time. We're going to have to develop a set of rules for the frequency of data capture, and it probably won't be the same for all things depending on the type and frequency of changes.

We have not yet really started writing the rules, and there's going to have to be a lot of discussion about that.

V.7 **QUESTION:** How will GPO offer with NARA on this project? NARA already holds large amounts of government information. Can GPO use this as part of the legacy collection?

RESPONSE: We have been delighted at the conversations we have had with NARA about all of these issues in terms of their willingness, and even eagerness, to be a partner with GPO in this. Many of you may realize that we turn over to NARA copies of everything we distribute to the depositories, and they keep that collection not as proof of anything the agencies ever did in publishing that information but as proof that GPO distributed it to the depository program. That's the construct that keeps that material in NARA and that's why that material was there when many of the ephemeral materials would otherwise have been discarded, if the agencies themselves had deposited them as part of their official records.

In their conversations with GPO, NARA has made it clear that they want their collection to be as comprehensive as possible. They were delighted with the idea that the ILS and the national bibliography would give them a way to inventory their collection and fill in missing pieces.

V.8 **QUESTION:** What if the Gates Foundation would like to digitize government documentss material? Would they be able to add to the collection assuming they agreed to PPA?

RESPONSE: I don't think that issue has come up. We are not adverse to partnering with other people as long as the terms for access are such that the content is in the public domain.

V.9 **QUESTION:** Will the project consider using other sources like Heinz Federal Register or LLMC's Congressional Record as an alternative to rescanning?

RESPONSE: We don't know those products enough to speak to them. But, we're open to looking at any kinds of partnerships that help us make this information available, as long as it meets the assumption that the content is available for no fee public access.

V.10 **QUESTION:** Will preservation masters or access files be available to information vendors for enhanced products free?

RESPONSE: Information vendors currently can get anything free that they can download on GPO Access or pull out of the archive through the PURLs or otherwise, the same as anybody else, and use it any way they wish.

We do have a program whereby they pay a fee for the source data files with embedded tag and typesetting codes for, say, the Federal Register, Congressional Record and so forth. That analogy could carry forward, so that they will have access to anything that's

accessible to the public for free, but there might be other kinds of things that they might want for which there would be a fee.

V.11 **QUESTION:** Whatever happened to the ad hoc committee report on the five-year retention rule?

RESPONSE: We had asked that committee to do some very specific things to be available to us before the American Library Association conventions last year so that we could address what changes might be needed in our policies for withdrawal and disposal of materials, if in fact the legislation passed that rescinded the five-year rule.

We didn't get the report until after those conventions. By then the appropriations bills which were the vehicles for that change were very far along and we had already stricken it. We did not make any attempt to put it back in. So we still have the report. But it became moot because of what was going on.

Over time the issue of eliminating the five-year rule becomes less and less significant. 86 percent of the titles are electronic, and are up on GPO Access. And so it doesn't become quite as much of an issue.

We really have not had time to go back and revisit the issue in terms of whether we're going to further revise the guidelines or further attempt to remove the five-year rule or whether we're just going to let it be overtaken by events.

V.12 **QUESTION:** On the OCR issues: when referring to a digital preservation master, do you mean only an image title or the image title to the OCR, with added text in some single or linked package?

RESPONSE: What was recommended in the meeting on digital preservation was that we preserve the individual TIFF files that represent each page that was scanned, that we do OCR, that we preserve the uncorrected OCR, that we then correct the OCR, and that the access files would include both a derived image of a page and the corrected OCR text so that it would be searchable.

V.13 **QUESTION:** Will there be both an image and an OCR master?

RESPONSE: In effect, there are both, but they are consolidated in the access file. So if you think of it in terms of a searchable PDF you may see the image file but in many of the searchable PDFs you can then search the OCR image and highlight in the image where that word appears so the two things have been integrated.

V.14 **QUESTION:** Isn't OCR’d text one of the first derivatives to be struck?

RESPONSE: OCR is one of the first derivatives to be struck. In a sense it is a derivative because you are scanning the image to create the OCR. But in another sense it is a separate file initially.

V.15 QUESTION: For certain types of textual information, i.e., statistical tables, OCR is not sufficient to insure access to the information. That is, structured data must be treated as structural data in this initiative. Seemingly just operational issues, but I feel they have a larger ramification as these assumptions are drafted as a genuine plan.

RESPONSE: That was in fact one of the issues that we raised at the meeting of experts on digitization. There is so much tabular material and tab tables of all kinds, statistical tables, appropriations and other financial tables, other tables that were just providing data elements of various sorts but in tabular form. One of the things that we need to do is to be able to reconstruct those as tables.

We have had an opportunity to review a project at Yale where they're converting statistical data back into tables rather than just into raw OCR. In fact, what they found was they had a much higher accuracy rate on the OCR of the numbers than they had on the OCR text, which I thought was kind of interesting. They did some check sums on columns that had numbers and found in most instances when the check sums didn't work it was actually because there was an error in the printed document, not because of an error on the OCR.

But what we have identified absolutely as an enhancement to the scanned data is to have those tables reconstructed through some sort of XML markup that would allow them to be used as tables and potentially allow the data to be manipulated. It's not very often somebody searches for a number in a table. But there is a need to understand the conjunction of a row and a column, to know that a data element exists that answers these two points of data, and how we might do that as we move forward with this tabular material.

VI. AUDIENCE QUESTIONS ADDRESSED AFTER THE MEETING

VI.1 QUESTION: What criteria will GPO use to determine priorities for digitization (besides the ranked list generated from the recent survey)?

RESPONSE: GPO is committed to digitizing the materials in the historical collections in the FDLP. In December 2004, Bruce James presented to Congress GPO's *Strategic Vision of the 21st Century*, which includes the objective of providing "access to all past, present and future Federal documents in a digital form that can be searched, downloaded and printed over the Internet at no charge." In addition to the results of the initial survey, GPO will consider the availability and condition of the documents, as well as the need to simultaneously process material of similar size and format to achieve production efficiencies. There will be additional opportunities for input from the library community as GPO continues to develop plans for the digitization of the legacy collection.