UC Federal Documents Archive Project Update

May 4, 2015
Elizabeth Dupuis, Project Lead

This update summarizes progress on the UC FedDocArc project since December 2014.

Selective Housing Agreement

In April 2015 we filed a final version of the Selective Housing Agreement, signed by all members of Council of University Librarians, to the Government Publishing Office and the California State Library, our regional depository. Confirmation of this document was a critical step toward assuring all parties of our commitment to first create a shared print copy, and then pursue access to a complementary digital copy.

Continued Investigation of Proposed Disposition Database

For US federal documents there is a well-recognized "needs and offers" process that allows depository libraries to make excess titles available to other depository libraries before withdrawing them. It is important that we find a process that is effective for geographically dispersed and large-scale collections. Last fall the disposition lightning team identified the open source software developed by ASERL for the disposition of federal documents as a model for further exploration. In January members of the FedDocArc Implementation Oversight Team viewed the "Demonstration of the ASERL Documents Disposition Database" webinar and talked with ASERL leaders about obtaining the software for a UC installation. In March we received access to the files although there is minimal documentation so our progress has been slower than we hoped. We continue to pursue this direction, as it appears to be the most streamlined and lightweight process.

Print Archive Disclosure

As recommended in the original report, we plan to disclose the print archive according to OCLC metadata standards. A draft of the standard exists and the RLFs have started to explore its implementation for the print resources. This work continues and is dependent on developments at OCLC around shared print monograph disclosure and registration.

Approved Disposition of Dis-bound and Scanned Items

After one copy of a title is adopted into the print archive, we will offer another copy to Google for sheet-fed scanning to fill gaps in the digital holdings available. These items are dis-bound before scanning so they can lay flat on the scanner and produce a better quality image. After the process, the single pages are impractical for use or re-shelving. In April our regional depository library confirmed approval for us to withdraw and dispose of the dis-bound copies that have gone through the sheet-fed digitization process. This is wonderful news and saves us from spending resources to have those single sheets banded, boxed, shipped back, and stored indefinitely.

Google Sheet-fed Digitization Pilot

We are continuing the pilot sheet-fed digitization project between Google and Berkeley that began in January 2015. Each month approximately 500 items which have a copy already stored at the RLFs are being digitized in this process, adding to the corpus of US federal documents available via HathiTrust. Google confirmed that they do color scan sheet-fed items. Through this pilot we gained knowledge of how long it takes to pull and process each batch of items (roughly 1.5 minutes to pull from shelves and another 1.5 minutes for list

prep work and packing onto Google trucks). We also have developed a better understanding of factors that make an item 'unscannable' for Google – such as too tight binding, brochures, and oversize foldouts. For these we are investigating other options. Our initial lists of eligible items came from Google; in the future we plan to contribute items from the RLFs. We have also agreed to work with the committee developing the registry of US government documents for HathiTrust on an approach to filling gaps they identify in their digital holdings. We developed criteria for items not recommended for destructive scanning which will be valuable once we reach Phase Three that moves to other campuses' collections; these items would be possible to digitize through page-turn scanning.

Enabling Full Text Access in HathiTrust

As recommended by the original UC Federal Documents Archive Project Team, we will rely on HathiTrust as the digital archive. We asked HathiTrust for the metadata signals they use to designate items as US federal documents and we are ensuring that the records we pass to Google for the items we send for digitization, and are sent to HathiTrust for ingest via CDL's Zephir, meet these criteria. Ensuring these markers are present enable users to get full text.

Metadata Comparison of RLF Records

With shared commitment of library staff at UCB/NRLF and UCLA/SRLF we gathered the records for items at the RLFs that met criteria indicating they are US federal documents. After much work to massage the data into compatible forms, we were able to create a master MySQL database from which we can now generate lists and track the state of review of each title. Currently we plan to maintain this database until we have reviewed and handled all titles at the RLFs; at that point we intend to replicate this process to catch any new federal documents that have been added to the RLFs since our original pull in December 2014.

RLF Shelf Check Sample

As recommended by the original UC Federal Documents Project Team, we sought to check how reliable our metadata records in the ILS are regarding the status of our government documents shelved at the RLFs. A sample set of 1000 items was generated following the ratio of US federal documents at each RLF (~63% SRLF/ 37% NRLF). SRLF and NRLF staff checked their items over winter break 2014, finding the RLF record status not reflecting the actual status of the item for only 4 items, representing 99.996% accuracy. The shelf check and problem solving work took an average of 1.5 minutes per item. A report is forthcoming to document an overview of the methodology, research process, quantitative analysis, findings, and staff time to undertake.

RLF Monograph Record Comparisons

From the master database of RLF government documents, we generated lists of records of that indicate exact matched copies are held at both NRLF and SRLF. Through this process we learned to separate out special collections (since they are not considered to meet our accessibility goals) and non-print formats (microfiche, microfilm, CD-ROMs, etc) as much as possible. One file is of OCLC-matched monographs (aka single volume monographs) of about 18K titles, and one file is of OCLC-matched multi-volume monographs (since campuses designate these differently, some titles are considered serials by some libraries) about 3.2K titles. The Implementation Oversight Team agreed to designate the NRLF as the retaining RLF in cases where there were duplicate copies at both RLFs; SRLF would gain the majority of freed shelf space through this process. From our work to date we have

developed a standard set of codes for decisions and guidelines to enable us to enlist the help of other staff in future reviews while ensuring consistency in decisions.

- Of the 18K single-volume monographs reviewed, ~ 113 items need special attention, meaning approximately 18K volumes will be adopted into the print archive and approximately 18K volumes duplicates will be removed from the RLFs for digitization or other disposition.
- Of the 3.2K multi-volume monographs reviewed, ~725 items need special attention. Each of these titles represents between two and a couple hundred volumes so estimating the total number volumes from this set is not possible at this point but this set represents a much larger total than it may appear at first glance.

Mini-Trial of Processes

Members of the Implementation Oversight Team and other key experts discussed the steps for all the key parts of this complex process – including file naming conventions, manage and update the master database, generate the lists, review and mark the lists, pull for other scanning, pull for Google scanning, returns and handle post-Google scanning, offer through disposition database, record changes for batch suppression/withdrawal, record changes for batch disclosure, check RLF shelves, recycle non-claimed items, and check and improve HathiTrust files. Berkeley will generate a small sample set (possibly 10 items) that we will step through all phases to further refine the details. After that we will work with larger batches to handle all items already reviewed and awaiting action. We intend to capture data about time spent on various steps to assist in estimating costs of future phases that will extend to the individual campuses.

ERIC Microfiche Adopted into FedDocArc

In Fall 2014 the Council of University Librarians endorsed the transfer of Berkeley's complete set of ERIC microfiche documents to the NRLF to become part of the UC Federal Documents Archive. Other UC Libraries have stored incomplete sets and it was determined to be most efficient to adopt the Berkeley set into the UC Federal Documents Archive in one step, rather than compare and consolidate holdings. In doing this, all other UC Libraries would be able to withdraw their holdings if they no longer wished to keep them. Berkeley will provide cataloging for all items to allow search by ED number and include URLs for e-versions when available. Requests for items from the ERIC microfiche collection come with specific document identifiers that would allow NRLF staff to pull the precise document, create a digital scan of the item, and fulfill the request as they do for journal articles. The collection moves to NRLF in June 2015 and records will be accessible around the same time.

Possible Ingest of Other Federal Document Digital Collections

The US Geological Survey (USGS) provides high quality digital scans for many of their publications on their web site. Rather than create new scans of these items – many with multiple, large sized foldouts – we are exploring a partnership between USGS, HathiTrust, and UC to ingest these materials to create even greater visibility and access as part of the digital archive we are building. The HathiTrust Board is considering hiring a dedicated project manager for their federal documents project; if that is approved they would help lead the policy and format evaluation work required and would eliminate the need for us to create duplicate scans of these items.

Confirmation with UC Libraries Contacts
In April each UC Library was asked to 1) confirm their agreement to have the disposition process handled centrally at the RLFs, 2) identify their preferred contacts for policies and procedures, and 3) identify their preferred contacts for metadata and records processing.

Many Hands
Special thanks during this period to: Jesse Silva, Elizabeth Cowell, Dave Rez, Lynne Grigsby, Emily Stambaugh, Erik Mitchell, Colleen Carlton, Tin Tran, Andy Kohler, Charlotte Rubens, Jeff Loo, Paul Fogel, Renata Ewing, James Mabe, Glenn Gillespie, Tamara Takeshita, and the staffs at SRLF and NRLF.