Against the Grain

Volume 29 | Issue 6 Article 9

December 2017

Federal Documents Archive-A Model for Preserving and Providing Access to U.S. Documents at the University of California

Jesse Silva
University of California Berkeley, jsilva@library.berkeley.edu

Federal Documents Archive: A Model for Preserving and Providing Access to U.S. Documents at The University of California

by **Jesse Silva** (Federal Documents/Political Science Librarian, University of California Berkeley) <jsilva@library.berkeley.edu>

United States government documents in print seem to have received a bad rap within libraries over the years, and not without some justification. They are cumbersome. They take up a lot of space. They are strange to many librarians, not just because of the additional rules and procedures, but also because of the weird Superintendent of Documents classification and call number system. While some may complain about the upkeep necessary for access to the printed documents collection, there is no denying that the information inside these materials has great research value and is vital to university collections as authoritative, primary sources used by many academic disciplines. With few caveats, these collections are also free from copyright, which can be a boon for those engaged in mass digitization projects.

The University of California (UC) is taking a unique approach to preserving U.S. government documents while also contributing to a mass digitization effort to provide access to a wider audience. **UC** has a large library system across ten campuses, holding over 39 million items, with numerous copies of federal documents. While **UC Libraries** are administratively separate, we share infrastructure to support our consortial work, including shared off-site shelving facilities, shared interlibrary borrowing agreements, shared policies and procedures, and a shared mission. In 2010, **UC Libraries** began examining our U.S. government collections to determine how to develop a shared collection across the system that would benefit all. In 2014, the conversation turned into action and the result is the **Federal Documents Archive (FedDocArc)**. This brief article will provide an overview of the **FedDocArc** project, highlight some milestones, discuss a few lessons learned thus far, and describe our next steps.

FedDocArc is an immense undertaking among the **UC Libraries**: the ten campuses have committed to retaining one shared print copy and one shared digital copy, via **HathiTrust**, of every U.S. document we hold in our collection. Materials that duplicate this shared collection may be deaccessioned from individual campus holdings or retained based on local need. Deaccessioned documents may be destructively digitized if not already in **HathiTrust** or offered to other depository libraries per current **Federal Depository Library Program (FDLP)** disposition requirements. If **UC** holds a single copy of a document, that single copy will be non-destructively digitized for **HathiTrust** and the print copy will become part of the physical archive. Titles deemed part of the print archive may be housed at one of the ten campuses or at one of the two off-site shelving facilities. The **FedDocArc** project is coordinated by a team from across the **UC** system, with a small core group guiding the project and individual campus members rotating on and off as their campus goes through the process of reviewing titles for inclusion.

UC began **FedDocArc** by reviewing the collections held at its Regional Library Facilities (RLF), which are off-site shelving facilities, one in the north and one in the south of the state. Over 150,000 titles were compared and reviewed for inclusion, with one copy being retained, and a second copy, if identified, being digitized or offered. Needless to say, monographs were pretty easy to work through this process. Serials and serial-like items such as multivolume monographs were another matter. In addition to the sheer number of volume and issue matching that needed to be done at the RLFs, **FedDocArc** revealed a previously unknown difference in cataloging practices between the campuses. Some campuses notated a serial title by date, while others notated the same serial title by volume and issue. "Problematic" is one way to describe the matching issues we encountered. Typical government cataloging issues were also discovered. For example, titles like "Report" with little other information in the bibliographic record were identified. Each of these problems required some hands-on work with the individual documents to resolve the match and make inclusion decisions. A project to provide cataloging enhancements, such as describing documents in ways that are more specific than just "Report," is being planned.

FedDocArc prompted **UC Berkeley** to hire a programmer to build an internal database for use in resolving the serials issues mentioned above. Still in development at the time of this writing, this database will help us track decisions so we do not discard materials we need to retain for the print archive. It will also help us identify potential gaps in serial holdings so we can target acquisition efforts to fill those gaps because **UC** is committed to creating a complete print archive of U.S. documents.

Milestones Thus Far

In the three years **UC** has worked on **FedDocArc**, we have reached a number of milestones. We are using the shared print practice of "disclosure" for documents that have been marked for inclusion in the archive. Disclosure is a notation in the bibliographic record that identifies the title as being part of a shared print archive. At the time of this writing, 194,080 monographic documents held at the off-site shelving facilities are in the process of being deemed part of **FedDocArc**. The southern off-site shelving facility has already offered 18,124 documents through the **FDLP** disposition of materials process. We will soon be training campus library staff on how to identify these documents in order to make local collection decisions. As official **OCLC** standards do not exist at this time, **UC** created a standard for disclosing monographs so we could continue our work. When the official **OCLC** standards are created, we have ensured easy identification of these documents so we can change them to meet the new standard if needed.

The first campus to undergo full collection analysis and review, **UC Riverside**, has nearly completed its review of monographic documents. **UCR** is contributing 25,662 documents to the digital archive, and plans to offer several thousand print documents to **FDLP** libraries in California. After the database is finalized, **UCR** will begin reviewing its serials while **UC San Diego** begins reviewing its monographs. As we continue adopting publications into the archive, **UC** will also produce a list of titles available in both the print archive and **HathiTrust** so campuses can make local retention decisions with the goal of decreasing the work future campuses need to do when they go through the **FedDocArc** process. Under **FedDocArc**, the titles a campus chooses to retain will be a local decision based on local needs. Libraries are not required to discard publications.

Why HathiTrust?

The **UC Libraries** became a member of **HathiTrust** shortly after its inception, and **UC** has contributed millions of print titles for digitization. **UC** has a number of mechanisms in place to support a continued partnership with **HathiTrust**, including its representation on the Board of Governors, so **HathiTrust** was a seemingly easy choice for a digital repository partner; however, **HathiTrust** is not without problems. One issue that **UC** is working to rectify has to do with foldouts, such as maps, large data tables, and other large format items. For example, in the original scanning efforts of libraries that contributed content to **HathiTrust**, foldout maps and charts from vital government titles, such as the **USDA** *Soil Surveys* and the **USGS** *Professional Papers* series, were skipped while the text was scanned. Based on the high technology costs to digitize foldouts at the time, this decision left gaps in the digital corpus which researchers and other users need filled for their work. As part of the creation of **FedDocArc**, **UC** is planning to rescan many

continued on page 27

Federal Documents Archive ...
from page 26

of these titles in-house, including the foldouts, and any other large format titles. These scans will also be contributed to **HathiTrust**.

Another issue is public access to digitized content. As of this writing, **HathiTrust** continues to only allow the download of full-text items in the public domain, including U.S. government documents, as a member benefit. The public still has the ability to search across full-text documents, so finding that obscure quote from a hearing or report is possible. The public can view and download a document page by page, not the entire document as one file, which can be a major frustration if a user who is unaffiliated with any member institution needs a lengthy document. While this access model is a great improvement over no access at all, it is not what **UC** considers full access. The single-page download restriction prevents members of the public from fully engaging with any of the federal documents contained within the database, including the thousands **UC** and other libraries have and will continue to contribute. **UC** firmly believes that fully opening up federal documents to the public without restrictions aligns perfectly with **HathiTrust's** mission to "contribute to...the common good by collaboratively collecting, organizing, preserving, communicating, and sharing the record of human knowledge."[1] Being able to share these digitized documents openly and without restriction would also fulfill the **UC Libraries'** mission to "provid[e] the broadest access to the world's knowledge."[2] **UC** is committed to working with and encouraging **HathiTrust** to remove the public download restrictions placed on federal documents, and we invite other **HathiTrust** members to do the same.

Working with **HathiTrust** has also been a great opportunity to brainstorm on various issues. **UC** and **HathiTrust** have been able to work through some of the issues both projects have encountered, such as reconciling various cataloging practices mentioned above. We have had preliminary discussions on resolving serials matching issues and identifying gaps in **HathiTrust**. A small example of gap filling: **UC** contributed several missing volumes of the *Statistical Abstract of the United States,* volumes that were non-destructively digitized so we may continue to retain the print for **FedDocArc**. **UC** and other digitization partners are also identifying publications we can target as priorities for digitization and inclusion in **HathiTrust**, such as titles from the **FDLP's** Essential Titles List.

In addition to **HathiTrust**, this undertaking has involved a number of players outside the University. **UC** signed a Shared Housing Agreement memorandum of understanding with the **U.S. Government Publishing Office (GPO)** in which provision of continued public access to the documents is explicitly spelled out. The **California State Library**, which oversees the **FDLP** in California, has been very supportive of the work we are doing to create a full collection of documents within the state. The **State Library** has allowed us some much-needed flexibility within the governing authority of the **FDLP**, so that we can work more efficiently to create the archive.

Next Steps

FedDocArc also requires **UC** to begin developing strategies to address several other issues. **UC** government documents librarians need to make some major decisions as **FedDocArc** moves forward. There are a number of questions to settle, such as which campuses will be responsible for contributing print publications to the archive and which campuses will contribute copies for digitization in the future. Will campuses split up the responsibility by agency, subject matter, or individual publications, or based on another option that has not yet been identified? What about CD-ROMs and other electronic materials, and born-digital content: how will these be included in **FedDocArc**? These are some examples of the questions remaining and the ongoing dialogue **UC Libraries** will need to continue among ourselves to resolve these issues.

The University of California Libraries are committed to completing the **FedDocArc** project and it has a great deal of support within the system. Having the collection digitized will open new avenues of discovery and research in scale and scope that had previously been unimaginable. **FedDocArc** has the potential to allow **UC** to open its collections to a large population outside the university, providing great public benefit, while at the same time retaining an archive of the print documents that will be preserved. **FedDocArc** is a project unlike anything the **UC Libraries** have done in the past, and we are looking forward to being able to share much of our collection of federal documents with the State of California, the nation, and the world via **HathiTrust**.

Endnotes

1. **HathiTrust**, "Mission and Goals," accessed August 31, 2017, *https://www.hathitrust.org/mission_goals*.
2. **The University of California Libraries**, "Vision and Priorities: UC Libraries," accessed August 31, 2017, *http://libraries.universityofcalifornia.edu/about/vision-and-priorities*.

End of Term 2016 Presidential Web Archive

by **Mark E. Phillips** (Associate Dean for Digital Libraries, the University of North Texas) <Mark.Phillips@unt.edu>

and **Kristy K. Phillips** (University of North Texas) <kristy.phillips@unt.edu>

Introduction

During every Presidential election in the United States since 2008, a group of librarians, archivists, and technologists representing institutions across the nation can be found hard at work, preserving the federal web domain and documenting the changes that occur online during the transition.

Anecdotally, evidence exists that the data available on the federal web changes after each election cycle, either as a new president takes office, or when an incumbent president changes messages during the transition into a new term of office. Until 2004, nothing had been done to document this change. Originally, the **National Archives and Records Administration (NARA)** conducted the first large-scale capture of the federal web at the end of **George W. Bush's** first term in office in 2004 (*https://www.webharvest.gov/*). This is noteworthy because, while institutions like the **Library of Congress**, the **Government Publishing Office**, and **NARA** itself have web archiving as part of their imperative, none of their mandates are so broad as to cover the capture and preservation of the entirety of the federal web. On April 15, 2008, **NARA** released the document "National Archives and Records Administration Web Harvesting Background Information," which detailed the reasons why the organization decided not to continue this large-scale archival practice during the following election in 2008. As such, a group of interested organizations gathered together to continue the project.

The End of Term (EOT) projects began with the **Internet Archive**, the **Library of Congress**, the **University of North Texas**, the **California Digital Library**, and the **U.S. Government Publishing Office** working together to fill the void left by **NARA** and archive the entirety of the federal web during the transition period in the wake of the 2008 presidential election. Since that first capture, new partners have joined the team, including **Harvard University** in 2012, and **George Washington University** and **Stanford University** in 2016.

continued on page 28