## Statement of Michael Furlough, Executive Director, HathiTrust
## Before the Committee on House Administration
## United States House of Representatives

## Transforming GPO for the 21st Century and Beyond: Part 3 – Federal Depository Library Program

**September 26, 2017**

Thank you for this opportunity to offer testimony before the House Committee on Administration on the topic of transforming GPO for the 21st century and beyond. My commentary will focus on the digitization of federal documents distributed by the Government Publishing Office (GPO) under the Federal Depository Library Program (FDLP). I will spend some time discussing the activities of my organization, HathiTrust, because I believe our experience will be useful as the committee considers how GPO should ensure access to and preservation of documents distributed through the program. My comments will pertain largely to academic and research libraries. Many depository libraries are not academic libraries, but most of the digitization of federal documents has occurred in these types of libraries. I request that my full statement be included in the written record of the hearing.

It is welcome news that Congress is considering modernization of Title 44, chapter 19, concerning the Government Publishing Office and the Federal Depository Library Program (FDLP). The current FDLP, first codified in statute in 1962, has served the nation well, ensuring that government publications are distributed widely, made available to members of the public, and secured for future generations. Libraries have enthusiastically stepped up to these functions and their institutions have supported this critical role for our nation. However, fifty-five years later, too much has changed to allow the current statute to continue to stand without revision.

*Background: the transition from print to digital*

In 1962, information was disseminated very differently, but there was something of a revolution underway. It could be said that the Federal Depository Library Program expanded along with America's research libraries. After the second world war and during the cold war, American universities grew in number and in size. The GI Bill educated a generation of soldiers and universities expanded to welcome them. Billions of dollars in federal research funds flowed to these universities, and the research results were reported in an ever-increasing number of academic and research publications. Universities were building larger campuses and new libraries were built to accommodate increased numbers of students, faculty researchers, and the books and publications they needed. It was necessary at this time for libraries to have the space to expand their collections; these publications needed to be physically close to the users to ensure that they could be readily accessed and read.  The top research universities could pride themselves on their library and the collections were an incentive to recruit both students and faculty.

Under the new FDLP program, many of these libraries assumed responsibilities as a “selective” depository library, and received a subset of the items distributed by GPO, and a smaller number of libraries became “regionals” with the responsibility to collect and make accessible all of these items.  Regionals and selectives were distributed widely throughout the continental United States to ensure that citizens could, with relative degrees of ease, travel to these libraries to consult government information as needed.  Millions of government publications in the form of books, journals, technical reports, magazines, maps, pamphlets, flyers, and other tangible formats were added to these library collections, and the libraries created space to accommodate these materials in their stacks.  The program ensured that many duplicates were distributed around the country, improving access and hedging against the loss of some copies. The relationship between regionals and the selective depositories associated with them, ensured a network of access to these physical collections.

By the 1980s we were on the cusp of a new information revolution.  Government information had begun to be distributed in digital formats in small numbers on CDs; depository libraries acquired equipment and deployed infrastructure to provide access to these materials. Networked access, however, was limited, and these formats presented significant preservation challenges that could not at the time be fully addressed.  But the advent of the World Wide Web changed everything.  In 1994 GPO launched GPO Access  (now FDSys) to begin providing direct access to citizens (much of this access happened in libraries, of course).  Over the following two decades we have seen a gradual but dramatic decrease in the number of print materials distributed under FDLP.  Most current federal information is distributed and accessed digitally now.  Citizens have a right to it, and they expect to have it with only a few clicks.  They also have come to expect the affordances of digital content with more robust discovery systems and searchability.

*The era of mass digitization and large scale library cooperation*

But what about all of those tangible government documents from the past?  They are still valuable—they provide a window into the history of our government and our nation and can provide us with crucial information that in many cases is still relevant today. Furthermore, print materials still offer many conveniences and affordances that digital items do not.  For most users, however, convenience wins over other considerations.  Librarians often say that for many users “if it’s not digital, it doesn’t exist.”

Many millions of federal documents have been digitized and made available online.  GPO itself has worked with some depository libraries, such as the University of Florida and Boston Public Library, to include documents digitized by the libraries in FDSys, and has partnered with the

Library of Congress on projects to digitize the *Statutes at Large* and the *Congressional Record.* Since the 1990s some publishers have developed extensive products based upon digitized federal documents. These are usually focused on materials of the greatest interest to researchers, such as complete runs of the Congressional Record and Hearings, papers of the Executive Branch, and others. The best of these products provide extensive cataloging and indexing and allow users to easily obtain greater contextual information that helps them understand what they are viewing. Unfortunately, these products are not produced with the individual citizen in mind, but are instead intended for sale and licensing to better resourced libraries and their users, and their terms restrict access to only the employees and enrolled students of the university. Most citizens do not have access to these collections and these collections do not even attempt to comprehensively represent the entire universe of federal documents. They are no substitute for a federal depository library.

Partnerships among libraries, commercial firms, and not-for-profits have resulted in the digitization of and ready access to millions of federal documents from library collections. Early in this century Google began partnering with libraries to scan their collections and add them to Google search databases. Google partners were large libraries with extensive collections; an initial focus often targeted large subject areas that included publications of all types, including documents produced by the government. A number of libraries also began working with the Internet Archive to undertake large scale digitization, while others worked with Microsoft. Google's program, which continues today, has been by far the largest effort to digitize library collections. These mass digitization programs have not been wholly altruistic—certainly Google as a commercial entity added value to its suite of services through digitization of books—but they have served the public interest by expanding access to the record of human knowledge held in our nation's libraries. For their part, libraries have provided the necessary support to retrieve, prepare, and manage the flow of materials to digitization partners. These collaborative efforts have made it possible to reconsider how the public can access historical materials and for libraries to work together in new ways to ensure that the print record of our culture is sustained into the future. And these programs have implications for the future of the Depository Library Program and how depository libraries preserve and provide access to government information sitting in their stacks.

### *HathiTrust's Federal Documents Program*

I serve as the Executive Director of the HathiTrust, an organization formed to take advantage of the opportunities presented by mass digitization conducted by Google and others. We were founded in 2008 by the University of Michigan, Indiana University, the universities of the Big Ten Academic Alliance, and the University of California. Today over 130 academic and research libraries make up HathiTrust, including both academic institutions and major libraries such as the New York Public Library and the Library of Congress. Our digital library, aggregated from the digitized collections of our members, now holds nearly 16 million digitized items. 5.8 million items are openly available for anyone, anywhere in the world to view and read. We do not charge users for access to this historical collection, nor does a user have to be affiliated with a member library to gain basic access to the collections.

Our mission encompasses providing access to and preserving the collections of our member libraries. HathiTrust developed the technology infrastructure for that access and for preserving the digital content; this significant infrastructure investment would not be possible for any single institution, but collectively the members of the HathiTrust enable its reality. But we are more than a digital library. HathiTrust provides a way to harness our members' collective expertise to solve large scale shared problems and catalyze new solutions. We help our libraries serve their users, and we do so in a way that contributes to the common good.

HathiTrust's members have from the start had a strong interest in the digitization of federal documents, and we have made significant investments to identify them, find copies that can be digitized, get them scanned, make them accessible, and ensure that those scans are preserved into the future. HathiTrust's members include 128 members of the Federal Depository Library Program, seventeen of which are regional depositories. Although HathiTrust itself is not affiliated with GPO or the FDLP, HathiTrust helps the depository libraries in our membership amplify their mission of providing public access to government information within the policies and structures of the FDLP.

Our members have digitized and added to HathiTrust over 1 million US federal documents, and we know that there is a grateful community of users who find these materials valuable. Documents make up just over 7.5% of our total collection, but account for about 10% of the overall usage of HathiTrust's collection. Historical federal publications are frequently among the most heavily consulted items in our collection. For example, in 2016 our fifth most heavily used item was a 1978 report of the House Subcommittee on International Organizations of the Committee on International Relations, titled *Investigation of Korean-American Relations*[1].

*Linking print and digital preservation*

Access to digital information must be coupled with robust digital preservation efforts. HathiTrust operates a certified trusted digital repository and is part of a national Digital Preservation Network. Materials added to HathiTrust must meet strict standards, aligned with those recommended by the Library of Congress and the GPO, to ensure that we can faithfully deliver information well into the future as technologies evolve. Our systems and services are operated as co-investments with our institutional members and our governance processes ensure transparency and accountability, contributing to the sustainability of our organization.

[1] United States. Congress. House. Committee on International Relations. Subcommittee on International Organizations., . (1978). *Investigation of Korean-American relations: Report of the Subcommittee on International Organizations of the Committee on International Relations, U.S. House of Representatives, October 31, 1978.* Washington: U.S. Govt. Print. Off.. https://hdl.handle.net/2027/pur1.32754077064610

While we are committed to preservation of digital information, digitization alone is not sufficient to ensure the preservation of print materials. Nor would I argue that digital can totally supplant print in all cases. Digital access can in many cases suffice for a user, but we believe that historical printed information should be preserved, and we believe that digitization aids us in addressing the task. In addition to our focus on federal documents, HathiTrust has also launched a program to establish a distributed, shared print collection of materials that mirror the entirety of our digital collection. Under this Shared Print program, members commit to retaining, in suitable environments, specified print materials for a minimum of twenty-five years. We operate a registry of these commitments, which libraries can consult to see how their collections compare with what others have committed to keep. In our first phase of this program, fifty HathiTrust member libraries proposed to retain over 16 million copies of books that correspond to about 4.2 million book titles held in HathiTrust (about 60% of the entire book collection). In our first phase, we did not make any special efforts to target federal documents for commitment, but over 220,000 government publication titles were committed in the program.

Overall libraries have transitioned from collection-oriented missions to ones focused on services and provision of access to information wherever it can be found. The brand new libraries built in the 1960s, 70s, and 80s, designed to hold one to three million volumes, are now at capacity or beyond. For example, the University of Minnesota library, which serves as a regional depository for Minnesota, South Dakota, and Michigan, constructed a main library building in 1968 designed to hold 1.5 million volumes, including the federal documents collections. Today it holds over 3.4 million volumes, far exceeding its capacity and curtailing space for users and programs. Spectacular off site storage facilities can and have been built for many libraries, but space and funding for new construction of “book warehouses” is not as readily available as it was fifty years ago. And, as more information is transitioned to digital access, the physical collections are less frequently accessed. While it was once critical to extensively duplicate books and journals at libraries all across the country, it is less necessary to do so now. There are, and will always be, large libraries that have a special obligation to collect and retain extensive, rich collections of print and digital materials, but it is less necessary that all libraries do so to the same degree that they once did. This is true of government information as well.

HathiTrust’s efforts to develop a shared print collection can help libraries choose what to keep and what to withdraw. The presence of a digital surrogate, coupled with a commitment by other libraries to retain a physical copy, will allow librarians to make informed choices about their collections as they attempt to manage them in more efficient and economical ways. This is in turn enabled through the development of registries of print retention information that document the titles and copies that have been committed for retention and the libraries that hold them. Our program builds on many existing regional and consortia-based shared print projects that libraries have independently established to share and coordinate retention of legacy print collections. However, we seek to connect these many regional efforts to help establish a national approach to collection management and preservation.

*Continuing challenges for digitization of federal documents*

There are many challenges involved in mass digitization and the preservation of its results. An obvious one is cost. HathiTrust now includes over 5.5 billion digitized pages. If we had to digitize those pages all over again today at our own expense, at a going rate of $0.10/page, it would cost HathiTrust $550 million dollars just to scan the pages and produce the images. That does not include costs related to cataloging, taking the items off the shelf, moving them to a scanning site, and returning them to the library, which could add at least tens of millions more dollars. This is one reason why working with commercial partners who are willing to share the cost of digitization has been so important to libraries and benefitted the public. Once scanned, it is relatively inexpensive to preserve these digital copies, and by developing common infrastructure for preservation and access HathiTrust can take advantage of economies of scale to make it extremely cost effective. Digitization of collections, including federal documents is costly, but it can be done more efficiently through coordinated and intentional efforts.

Another well-known challenge is quality: if you digitize a lot of books, you inevitably generate errors. Over the last fifteen years scanning processes and quality correction have continued to improve, but in an endeavor of this scale, it may never be possible to fully eradicate all errors. Unfortunately, many problems originate not from the scanning process, but from physical flaws in the book itself, which may have been printed incorrectly or have been damaged since it was printed. With federal support from the Institute of Museum and Library Services, the University of Michigan undertook a research project in 2010-11 to sample digitized items in HathiTrust and categorize the nature and extent of errors found in the scans. This study confirmed the presence of errors which can interfere with readability of information, such as warped or blurry pages, or even missing pages. The study also noted that the vast majority of pages reviewed were not problematic. As digitization processes improve we will need to continue to monitor how quality improves over time. In digitization projects as large as these, we can accept some error as a price to be paid for improved access, but we also recognize the need to develop cost-effective quality control mechanisms that can detect and mitigate such errors. HathiTrust has now developed processes that allow users to report errors as they are found, which in turn initiates a process to fix or rescan the problematic copy. We are also investigating mechanisms to report to users what we know about the quality and completeness of the work to determine the suitability of a digital reproduction for a given purpose.

Yet another challenge is cataloging and metadata. Librarians have well defined standards for the description and documentation of the books in their collections, and these standards allow for local practices which are important for the users of that library. What digitization has taught us is that the proliferation of local practices, which were just fine at the local level, can create a huge problem when you collect millions of books: standards give way to highly diverse ad and sometimes conflicting cataloging records. These problems are especially acute for government documents. Prior to 1976 many federal publications were not consistently cataloged, and we have no definitive source that lists all publications of the federal government through time. We do not even know for certain how many government documents exist. While GPO's Catalog of U.S. Government Publications is a tremendous and reliable resource, GPO admits that it is not complete, especially for pre-1976 publications. We have no reliable complete record of the entirety of publications of the US government.

HathiTrust has attempted to address this problem through the development of a Federal Documents Registry. Our goal is to develop an inclusive and comprehensive database that captures known, extant federal documents, and to link this to information we have about the holdings of materials in library collections. The primary reason to do so has been to identify what federal documents have not been digitized, locate them, and get them scanned. However, we have found that the Registry has allowed us to identify over 300,000 federal documents that had already been digitized but which were incorrectly cataloged and not identified as documents. If not properly identified, they may not be viewable as public domain resources. We believe that the Registry could ultimately also support depository libraries and others in assessing the completeness and extent of federal documents collections.

Developing a complete list of all federal documents ever produced, and creating a comprehensive collection of digital federal publications are both activities with an ever-receding horizon. It is difficult to say when and if all historical federal documents will be available digitally and complemented with secure print copies for long term retention. But these challenges should continue to be addressed and the modernization of the Federal Depository Library Program can help to do so.

*Implications for revisions to Title 44 Chapter 19*

Free public access to government information is an essential part of our democracy and we believe that the Federal Depository Library Program should be enabled to continue to fulfill its critical mission. My focus has been primarily on the legacy collections of government documents, but Title 44 should also continue to ensure that public access to future digital publications can be guaranteed. However, Title 44 must be updated in ways that recognize how depository libraries and users of government information have changed and simplify how depository libraries can work together to fulfill the program's mandate.

Federal depository libraries have made significant investments in services, programs, and infrastructure to fulfill their obligations under the program while continuing to support the user communities that they primarily serve. Many of these libraries have partnered with Google, Internet Archive, HathiTrust, and others to develop large-scale solutions to the massive challenge of access to and preservation of government documents. They have done so largely independently of GPO, whose regulations and resource constraints have put them at odds with the innovation and creativity of its depository libraries. Over the past fifty years libraries have transformed to focus increasingly on service provision and digital access and less on the development of exclusive local collections. We continue to need a number of libraries to serve as regionals with the responsibility of maintaining physical collections. But in an era where the public and library users demand digital access that can be readily provided, it is not necessary to continue to enforce the levels of widespread print duplication and collection redundancy that the FDLP originally envisioned and continues to require.

Title 44 should support comprehensive digital access to future and retrospective government documents and should provide for measures that protect the privacy of users who access digital government documents.[2] Title 44 should provide mechanisms to ensure that depository libraries can work together to facilitate this work. These libraries should have the latitude to collaborate and ensure that the entire record of government publications is well documented, that an appropriate number of physical copies are committed for retention, and that these commitments are disclosed in ways that allow depository libraries to make withdrawal decisions. The future requires fewer print collections and greater emphasis on coordinated collection management and service provision. Preservation and access are the business of librarians, and they can be trusted to work together to ensure that public information endures.

An example of this type of collaborative effort can be found among the Association of Southeastern Research Libraries (ASERL), which has established a "centers of excellence" approach in which specific libraries have agreed to focus on cataloging and collecting efforts on specific agencies or sets of materials. However, this approach does not address the problem of extensive print redundancy and the costs that come with it. Another example would be a multi-state regional depository model such as that found at the University of Minnesota, which serves as the regional depository for its home state as well as Michigan and South Dakota. Minnesota has piloted projects that enable the selective depositories to discard duplicative materials and place them into a digitization workflow. This addresses cost challenges associated with managing federal documents collections and improves access to these documents.

GPO's FIPNet program is in some ways analogous to HathiTrust's shared print program, in that it provides for a limited number of depository libraries in separate census districts to retain print copies so that others can discard. But FIPNet should not be operated in isolation from the dozens of regional print retention programs that focus on books and serials other than federal documents, which are already underway and well established. A shared print program for federal documents should follow the best practices and methodologies established by these programs to identify materials for withdrawal and retention. Nor should GPO attempt to

[2] The American Library Association's Code of Ethics states that librarians "protect each library user's right to privacy and confidentiality with respect to information sought or received and resources consulted, borrowed, acquired or transmitted." American Library Association Code of Ethics: http://www.ala.org/tools/ethics.

duplicate commitments, registries, or other related data resources if they have already been made through other programs and advance the goals of the FIPNet.

While GPO has partnered with some depository libraries to digitize historical materials and add them to FDSys, GPO should avoid new digitization of any materials that have already been scanned unless it is absolutely necessary. In other words, if a collection has been digitized by a library or could be found at the Internet Archive or in HathiTrust, GPO should not seek to duplicate this work so that it could be included in FDSys. There are far too many documents that have not yet seen the digital light of day to duplicate these efforts unless they are critically necessary. Digitized versions should be recognized as effective access substitutes for print documents, and their existence should be factored into the process of deselection and retention for preservation and future access. These activities will require further development of cost-effective quality control systems to ensure that digitized materials reliably reproduce the information represented in the publication

GPO should be able to partner with libraries to identify investments of mutual benefit, such as quality control systems, better cataloging and discovery systems, and metadata remediation. It should be much simpler for GPO to work with its depository libraries and other potential partners. HathiTrust and GPO have previously discussed issues of common interest, but it can be challenging to provide metadata or digital scans to GPO because regulations require that GPO provide something of value in exchange. In other words, GPO should be empowered to accept metadata, digitized materials, resources, and/or services that support its programs without providing something of value in exchange. And they should be free to make non-financial contributions towards collective library efforts without requiring something of value in exchanges. Let me be clear that I am not proposing that libraries can and should do work without support from GPO or funding from outside sources. I have no expectation that Title 44 will be matched with significant new appropriations to fund digitization. However, where resources exist and can be made available, GPO should be free to draw on these and to contribute to common good efforts undertaken by their depository libraries

HathiTrust and its member libraries are eager to see the FDLP program thrive. But to thrive it must be enabled to operate in ways that promote cooperation and collaboration among depository libraries, support digital access, and reduce collection redundancies that impose significant cost burdens on its depositories. I believe that HathiTrust's work offers some useful models to consider and could help to advance many of GPOs stated goals. Thank you for providing the opportunity to offer my comments in support of legislative changes to Title 44.