2013-05

Federal Library Bibliographic Records Analysis: Initial Findings, Use Cases, and Recommendatio

Library of Congress, FEDLINK

Monterey, California. Naval Postgraduate School

Federal Library Bibliographic Records Analysis:



Initial Findings, Use Cases, and Recommendations

LIBRARY OF CONGRESS
FEDLINK

May 2013

About FEDLINK Research

This research was sponsored by FEDLINK as part of the FEDLINK Research Agenda. This report identifies significant issues within Collections Management, one of our six priorities for research. Research in this topic focuses on methods to improve the way in which we handle information. Through a series of forums, educational events, and research, FEDLINK provides information and insight into issues affecting the federal, as well as the broader, library and information science community. By working in collaboration with a wide range of stakeholders, FEDLINK provides opportunities to explore challenges facing our constantly changing information landscape and to develop a shared vision for library and information science research.

Acknowledgements

The following report is an interagency collaboration of federal library and information professionals who represent science, technology, engineering, and mathematics or medical (STEM) libraries. Over the past year a group of library and information science professionals have developed a pilot project which explores the federal STEM holdings. Through a combination of in-person meetings, virtual meetings and online exchanges, the participants co-authored and edited the following report. In writing this document, participants encourage future collaboration in this area.

Pilot Project Participants and Co-Authors

Carol Ayers, ***National Program Manager,*** National Forest Service Library

Deborah Balsamo, *National Program Manager,* EPA National Library Network

Julie Blankenburg, *Supervisory Librarian,* Forest Products Laboratory, National Forest Service Library

Sally Bosken, *Director,* U.S. Naval Observatory Library

Christopher Cole, *Manager, Business Development,* National Agricultural Library

Blane K. Dessy, *Executive Director, FEDLINK,* Library of Congress

Thomas Doughty, *Metadata Services Librarian,* Dudley Knox Library, U.S. Naval Postgraduate School

Stanley Elswick, *Database Librarian,* Library Information Services Division, National Oceanic & Atmospheric Administration

Michael Esman, *Chief Collection Development Librarian,* National Agricultural Library

Nancy Faget, *Librarian,* Army Research Laboratory

Mike Handy, Deputy Assoc. Librarian, Library Services - Programs, Library of Congress

Anne Harrison, Librarian/Network Program Specialist, FEDLINK, Library of Congress

Richard L. Huffine, Former U.S. Geological Survey Libraries Program, U.S. Geological Survey

Neal Kaske, *Chief, LISD,* Library Information Services Division, National Oceanic & Atmospheric Administration

Irena Kavalek, *Supervisory Librarian,* U.S. Geological Survey

Rosa Liu, Manager, Research Library & Information Program, NIST Research Library

Stephen Short, *Program Planning Specialist,* Library of Congress

Jamie Stevenson, *Head, FEDLINK Research,* Library of Congress

Amanda J. Wilson, *Director,* National Transportation Library

Note: Titles reflect participants' roles and organizations as they were in 2013 for the pilot project.

Table of Contents

Introduction

The Federal Library and Information Network (FEDLINK) at the Library of Congress is a consortium that serves all federal libraries and information centers worldwide. While FEDLINK works to centralize and streamline procurement, FEDLINK also serves as a forum, coordinating cooperative activities and services. Federal libraries have, by and large, independently managed their institutions' information resources. However, there are opportunities for collaboration. In 2012, the Office of Management and Budget (OMB) along with the General Services Administration (GSA) designated FEDLINK as the Executive Agent for Strategic Sourcing of Information Resources in the federal government. This designation from OMB and GSA, combined with the goals outlined in the 2012-2016 FEDLINK Business Plan, inspired FEDLINK to develop and implement a research agenda to:

1. Conduct research and report on issues and policies that affect the federal information community

2. Identify, prioritize, and recommend solutions to meet the challenges of providing information services to the federal government

To that end FEDLINK has identified six areas as priorities for research and is facilitating interagency collaboration to provide comprehensive information about the information landscape and identify opportunities for streamlining and sharing resources.

This pilot project is the first study within Collections Management and seeks to analyze the holdings of federal STEM libraries, specifically examining areas of overlap, duplication, and uniqueness of current holdings. This analysis may be leveraged by agencies more effectively managing their collection and developing collections and preservation policies.

Background

Federal libraries provide information services to vast constituencies who seek information across the spectrum of intellectual and creative endeavors. Collections in federal libraries reflect this broad information content and federal librarians now manage a large inventory of resources and data describing the collections as part of the library collections.

In the current environment, federal libraries are being asked to better leverage their collections, reduce their space requirements based on the shift from tangible print to electronic resources, identify cost savings where possible, and increase collaboration across the federal government.

Currently, there is not an established process for analyzing library resources across the federal government.

FEDLINK sponsored a research pilot project to better understand the requisite processes for analyzing bibliographic holdings and overlap among federal libraries. The sharing of Science, Technology, Engineering, and Mathematics (STEM) collections has been a topic of great interest in FEDLINK discussions about shared collection management. STEM collections were therefore the starting point for comparative analysis of federal library holdings.

The pilot was started in the summer of 2012 and included a small number of federal libraries with significant holdings in the STEM fields. Eleven libraries provided bibliographic records for use in this pilot project:

National Agricultural Library (http://www.nal.usda.gov/),
National Library of Medicine (http://www.nlm.nih.gov/),
Library of Congress (http://www.loc.gov),
U.S. Geological Survey (http://library.usgs.gov/),
U.S. Army Corps of Engineers (http://www.usace.army.mil/Library.aspx),
U.S. Forest Service (http://www.fs.fed.us/library/) ,
U.S. Forest Service – Forest Products Laboratory (http://www.fpl.fs.fed.us/products/library/),
U.S. Naval Observatory (http://aa.usno.navy.mil/library/),
U.S. Naval Postgraduate School (http://www.nps.edu/Library/),
Nuclear Regulatory Commission (http://www.nrc.gov/reading-rm.html), and
National Oceanic and Atmospheric Administration (http://www.lib.noaa.gov/).

In addition to comparing holdings across federal STEM libraries, we also compared federal holdings to the HathiTrust Digital Library. The HathiTrust was identified as a repository that was likely to contain a large number of digitized resources also held by federal libraries. The HathiTrust began in 2008, as a collaboration between thirteen universities, the Committee on Institutional Cooperation, the University of California system, and the University of Virginia. They established a repository to archive and share their digitized collections. HathiTrust bibliographic data was downloaded from their online collection.

Library collections in this initial pilot were limited to those materials represented by bibliographic records available for export from the holding institution's library catalog. The initial comparison was at the bibliographic record level, or title level. Specific volume holdings for serials and multi-volume sets were not addressed.

The results from the STEM collection analysis pilot and similar projects have the potential to impact management decisions ranging from collection coordination, shared repositories,

acquisition strategies, and consortium purchases. Because federal institutions may use this and similar data to determine policies and procedures, it is imperative to provide full documentation of the analysis and results.

Pilot Project Methodology

Holdings of STEM libraries used in this study include cataloging performed by the various libraries through 2011. To varying degrees, the records also include 2012 data. Almost all files were compressed to facilitate transfer. Data transfer occurred largely through File Transfer Protocol (FTP). Two of the larger datasets were transferred via flash drive.

Though all datasets for the current project were created with strict input guidelines in the form of MARC (MAchine-Readable Cataloging) and widely-adopted cataloging rules, they nevertheless contained considerable variance and inconsistencies among the libraries and even within a given library. Individual library staff and library policies change over time, so having federal library staff assistance in the analysis proved essential to isolating and explaining incongruities in the data. This variance made it necessary to treat the bibliographic records as free text, or text with high variance that is wrapped in a controlled structure. An in-depth explanation of data manipulation done to prepare the data for analysis can be found in the companion document to this report: "Federal Library Bibliographic Record Analysis: Processing Documentation."

Core Data and Duplication Identification

Four relatively unique identifiers offered common data points for comparison among the federal library records:

International Standard Serial Number (ISSN),
International Standard Book Number (ISBN),
Library of Congress Control Number (LCCN), and
OCLC number.

Initial results showed instances in which these identifiers were used for multiple items with different titles. These results were essentially false duplicates that, upon analysis, were for entirely different works.

To reduce false duplicates, the main title was used in conjunction with the unique identifiers to determine overlap. A given bibliographic record from one federal library matched a record in another federal library only when both a normalized control number (the ISSN, ISBN, LCCN, or OCLC number) and the normalized main title matched their counterparts in the respective records.

In addition, three elements common to all records are needed for most analysis: publication date, publication place, and publication language. The inconsistent use of these elements in the provided records precluded them for use in determining duplication. However, they facilitate grouping in broad categories for preliminary analysis, e.g. date ranges.

Data Layering and Normalization

All analysis of source data reported here relies on the concept of data layering. Data layering is the application of sequential or parallel data manipulation steps to raw source data in order to extract valid meaning. An important feature of data layering is the retention of the raw source data. Processing and manipulation of the source data occur virtually with no alteration of the source data.

The process of layering is critical to the overall validity of data analysis. With layering, it is possible to identify flaws in analysis through replication and deconstruction. Replication occurs by starting with the source data and systematically applying the steps applied. Deconstruction involves reversing data manipulation steps to revert back to the original source data. Both methods of verifying validity rely on source data that has not been altered.

The source MARC records contain many inconsistencies which impede analysis using the core elements. By using a combination of IBM BigSheets and desktop data manipulation tools, it is possible to correct some data problems and, at the very least, group unusable data as one, e.g., "undetermined." This helps generate consolidated counts of records by the core data elements such as date and place of publication.

Consider the need to show the number of records by country name. The raw data, in this example, contains country codes but no names. One option would be to replace the country code in the source data with a country name. This would, however, permanently alter the original source data. Layering makes it possible to achieve the transformation virtually without altering the source data.

The first example simplifies a more complex process that sometimes involves more than twenty layers to achieve the virtual result desired (see Example 1). A key aspect of this process is that the raw data, even the flawed country codes, are not changed and can be retrieved if needed.

The layering techniques were most critical when addressing inconsistencies in the unique identifiers. Consider, for instance, the difficulties encountered when working with ISBN's. The International Standard Book Number (ISBN) is a unique identification sequence for books. In 1975, the International Standards Organization (ISO) promulgated the ISBN as a replacement for the Standard Book Number (SBN) which was developed in 1965. The number of characters has varied over time. Initially, there were nine characters and currently there are 13 characters. Like the ISSN, the last character is a check digit and an 'X' appears if the check sum is 10.

Example 1: Layering to find counts by country name.

Base Layer = record id, country code

Virtual Layer 1: Join code to table containing name, or label
[country code + country name, record id]

Virtual Layer 2: Replace flawed country codes with 'Undetermined'
[If country code has no corresponding name, use 'Undetermined']

Virtual Layer 3: Show count by country
[Count (record id), country name or 'Undetermined']

The variations in length of the ISBN pose some problems. Some libraries have retroactively appended a '0' to the beginning of the nine-character original SBN's to make them consistent with the ISO's ten character ISBN. This practice was not uniform. In addition, newer ISBN's have a '979' or '978' prefix, making them 13 characters in length. The potential exists for some libraries to append this new prefix retroactively, just as had been done with the old SBN's.

The ISBN also has been input into bibliographic records in several different ways. For example, the fictitious ISBN "0664226612" could be manifested in a bibliographic record in several ways:

0664226612
0664226612 pbk
0664226612 (pbk)
0664226612 (pbk. : alk. paper)
978-0664226612
978-0-664-22661-2

The variances in length and input require several major data manipulation steps (see Example 2).

Through layering steps such as those in Examples 1 and 2, all unique identifiers and main titles were normalized to reduce variances that could preclude the identification of matches. This effort made it possible to compare data that would otherwise be incompatible.

Example 2: Layering to find counts by country name.

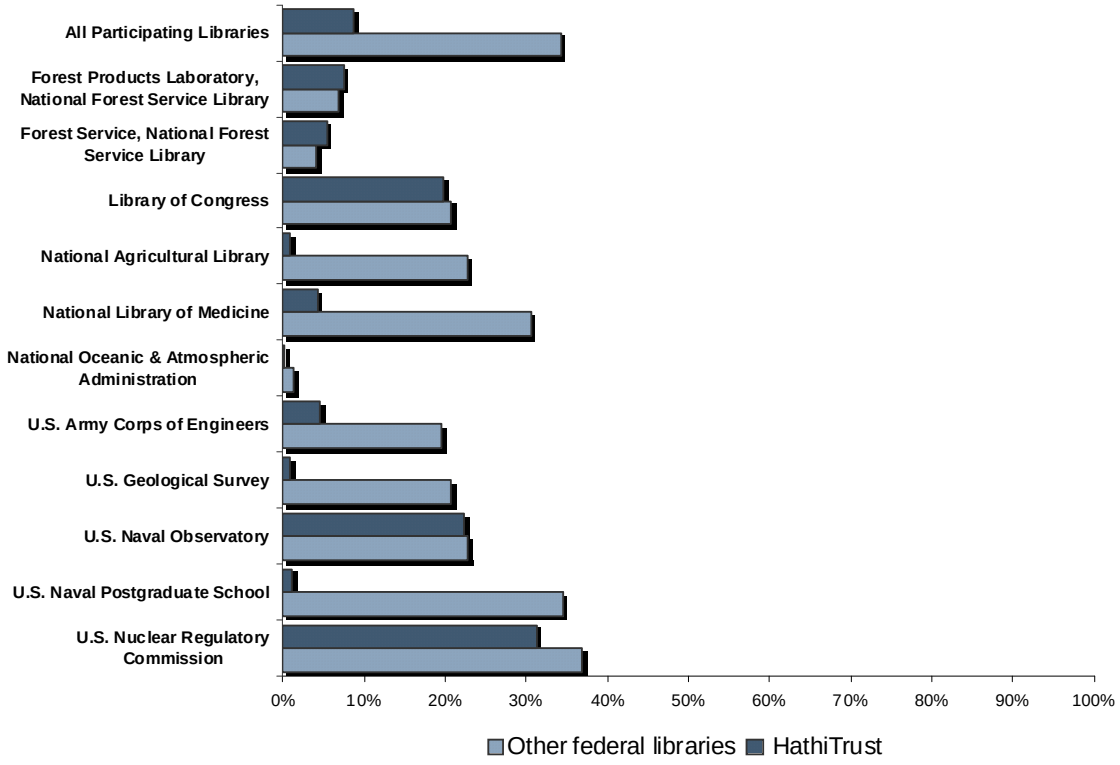
Base Layer: Limit to tag 020 |a

Virtual Layer 1:	Remove hyphens {Substitute(ISBN,'-',')}
Virtual Layer 2:	Remove leading white space {RIGHTTRIM(ISBN)}
Virtual Layer 3:	Limit to ISBN length
Virtual Layer 4:	IF ISBN starts with 978 or IF ISBN starts with 979 then {LEFT(ISBN,13)} ELSE {LEFT(ISBN,10)}
Virtual Layer 5:	Remove all leading and trailing white space {TRIM(ISBN)}
Virtual Layer 6:	Remove all characters EXCEPT 0-9 and (X or x)
Virtual Layer 7:	IF ISBN length is 9, add leading zero

Initial Findings

The comparison of bibliographic records shows that eight percent of the total federal library records examined matched a bibliographic record in another federal library. When comparing the federal library bibliographic records to HathiTrust bibliographic records, the total overlap found is lower, coming in at six percent (see Chart 1).

Chart 1. Percent of bibliographic records matching another library bibliographic records and HathiTrust bibliographic records



Overlap among Federal Libraries

The overlap among federal library holdings ranged from four percent – the Library of Congress matches to other libraries – to 37 percent, the U.S. Nuclear Regulatory Commission matches to

other federal libraries. The overlap reflects title matches and does not compare specific holdings information of multivolume sets, serials, or other holdings with multiple pieces. For most participants, the greatest duplication among collections can be found with the Library of Congress. A notable exception is the Forest Service data for which the greatest number of matches can be found when compared to the National Agricultural Library data (see Table 1 for details). Note that the two Forest Service libraries intentionally have duplication of Forest Service published materials; yet, they still do not have a large overlap. [1]

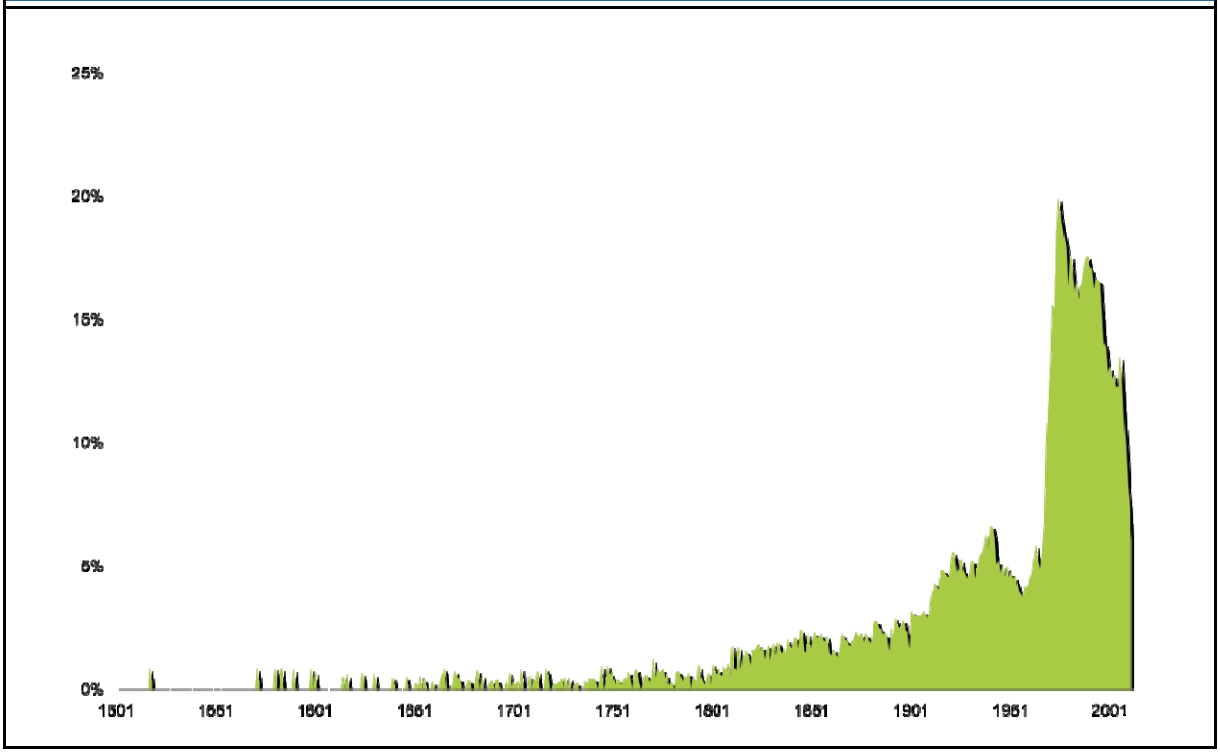
Table 1. Percent of total bibliographic records found at other institution

Library	Forest Products Laboratory, National Forest Service Library	Forest Service, National Forest Service Library	Library of Congress	National Agricultural Library	National Library of Medicine	National Oceanic & Atmospheric Administration	U.S. Army Corps of Engineers	U.S. Geological Survey	U.S. Naval Postgraduate School	U.S. Nuclear Regulatory Commission	U.S. Naval Observatory
Forest Products Laboratory National Forest Service Library	0%	21%	16%	15%	1%	2%	1%	2%	1%	1%	0%
Forest Service, National Forest Service Library	6%	0%	2%	3%	0%	0%	0%	0%	0%	0%	0%
Library of Congress	0%	0%	0%	1%	2%	0%	0%	0%	1%	0%	0%
National Agricultural Library	1%	1%	19%	0%	3%	2%	0%	1%	1%	0%	0%
National Library of Medicine	0%	0%	22%	3%	0%	0%	0%	0%	0%	0%	0%
National Oceanic & Atmospheric Administration	1%	0%	28%	7%	2%	0%	2%	6%	7%	1%	1%
U.S. Army Corps of Engineers	1%	0%	18%	4%	1%	3%	0%	4%	4%	2%	0%
U.S. Geological Survey	0%	0%	20%	4%	1%	4%	1%	0%	2%	1%	0%
U.S. Naval Postgraduate School	0%	0%	34%	2%	1%	4%	1%	2%	0%	1%	0%
U.S. Nuclear Regulatory Commission	1%	1%	33%	5%	3%	6%	4%	6%	9%	0%	1%
U.S. Naval Observatory	0%	0%	22%	1%	1%	3%	0%	2%	3%	1%	0%

[1] The data reflects the percentage of records in the holding institution (left column) found in another institution (vertical columns). The percentage is derived from taking the total holding institution (left column) records and dividing by the number of matches with another institution (indicated in vertical column).

The majority of the overlap among the federal libraries is for materials produced from about 1965 to the present day. The overlap appears to be declining (see Chart 2). However, the cause of the decline, or even whether it is a real decline, is not clear. It could reflect an arrearage in cataloging and processing of materials more than a reduction in acquisitions.

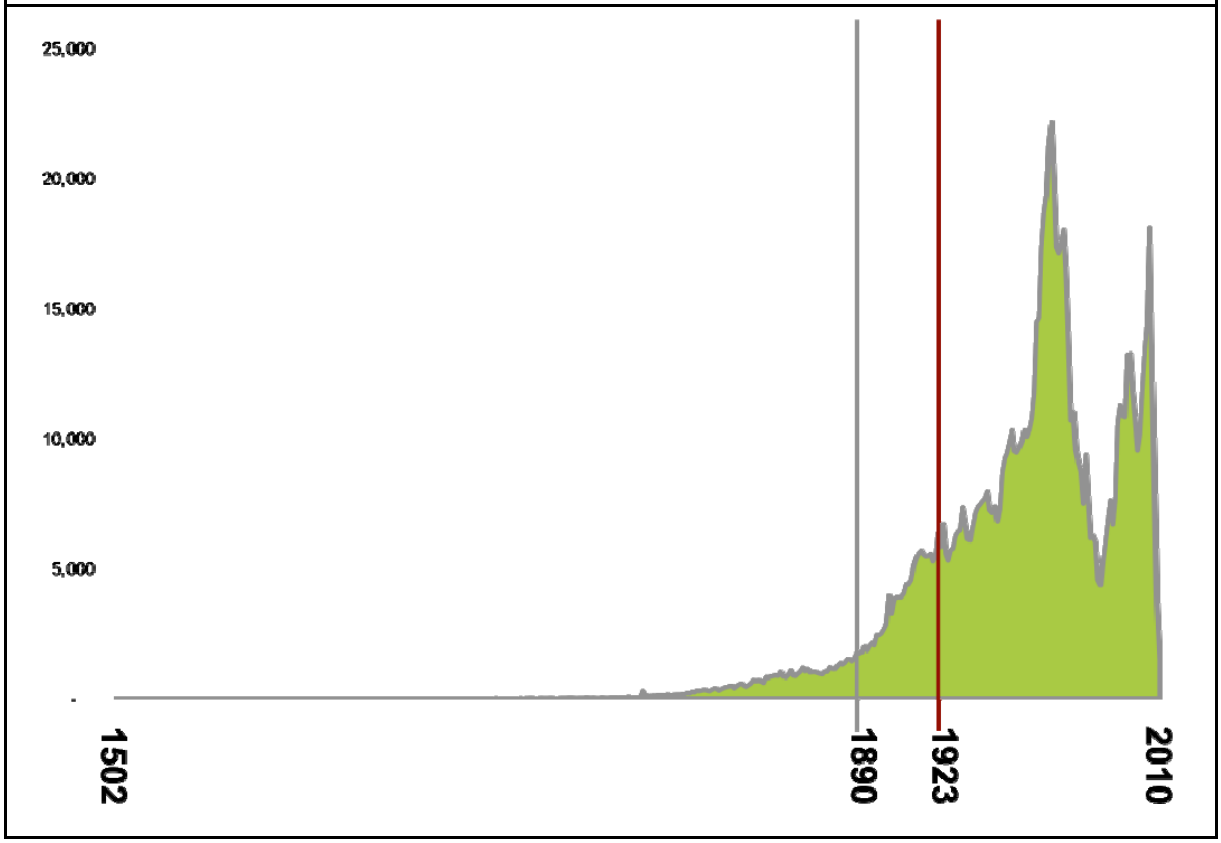
Chart 2: Federal library duplication by year



Overlap with the HathiTrust Digital Library

The overlap of the pilot libraries with HathiTrust ranges from one percent to twenty-two percent, or the two Forest Service libraries (1%) and the U.S. Naval Observatory Library (22%) respectively (see Chart 1 for details). As with the comparison among the federal libraries, the overlap reflects title matches and does not compare specific holdings information of multivolume sets, serials, or other holdings with multiple pieces.

Chart 3: Time distribution of federal record matches with HathiTrust records



The HathiTrust data includes a large number of materials that are not accessible because of copyright and legal restrictions. Only eighteen percent of the federal library matches to HathiTrust records were published prior to 1923. The analysis for this pilot does not include an examination of access and restrictions, but a large portion of the post-1923 material will likely be restricted because of U.S. copyright laws. Despite the HathiTrust having digital files for these materials, the legal status makes them unusable or inaccessible; thus as much as 82 percent of the federal library matches to HathiTrust data might not be usable.[2] If international copyright is taken into consideration, the number of potentially unusable files jumps to 94 percent given an 1890 cutoff year (see Chart 3 for time distribution of HathiTrust duplicates).

[2] Many of the duplicates represent Federal publications which have no restrictions so the total of unusable HathiTrust files should fall below 82 percent.

Implications of STEM Analysis

General implications for federal STEM libraries can be identified in a number of areas. The following areas are a starting point for conversation about services and functions that are common to federal STEM libraries.

Impact on Acquisitions and Collection Development

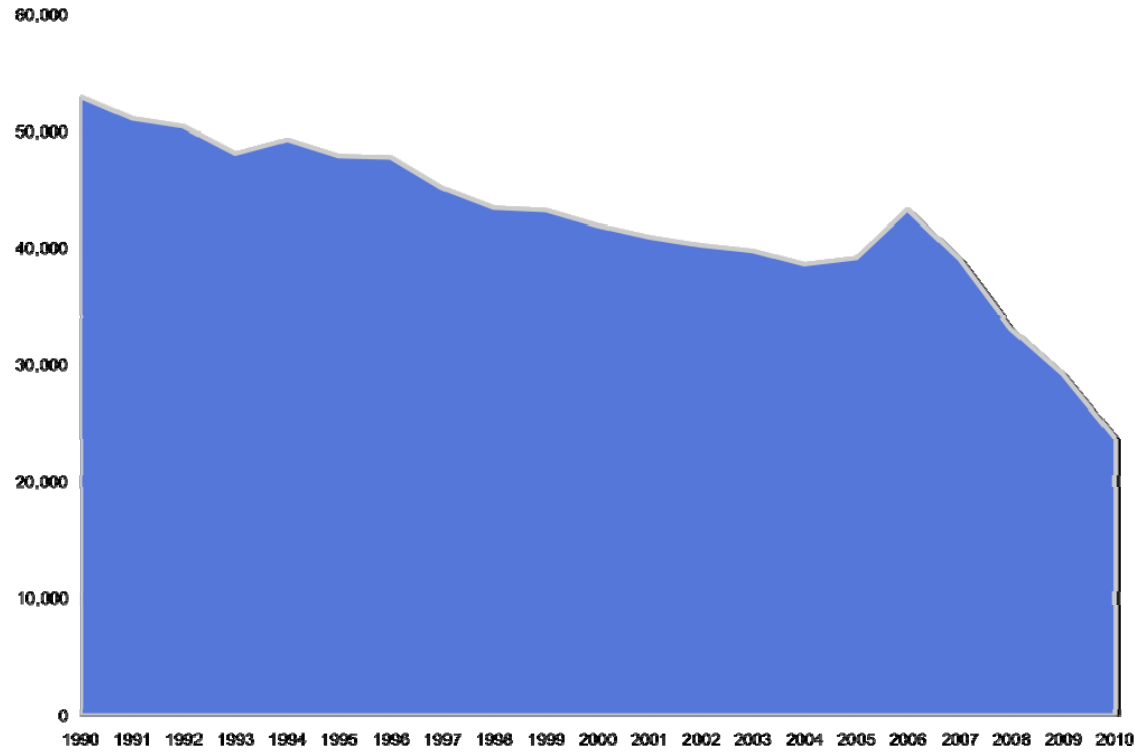
Federal libraries are diverse both in terms of size and subject specialization. They range from the large National Libraries (Library of Congress, National Library of Medicine, and National Agricultural Library) to medium size libraries, to one librarian/one room libraries. Most federal libraries may be considered as special libraries with the possible exception of the Library of Congress, which has a mandate to collect in most areas. The libraries participating in the STEM collection analysis have unique subject collecting responsibilities.

Over the past twenty years, 1990-2010, an average of 14 percent of cataloging records of one federal science and technology collection duplicated the same cataloging record of another federal science and technology collection in the same year (see Chart 2). In recent years that number has steadily declined from 11 percent in 2008, 9.5 percent in 2009, and eight percent in 2010. This drop in duplicate title acquisitions reflects materials budgets in a period of retrenchment. The data also suggests that libraries are acquiring materials that may not be available at other libraries.

During this twenty year period, annual cataloging record production among the surveyed libraries remained fairly stable at 287,000 to 327,000 records, with the duplicate record rate gradually dropping in recent years. It is unclear what accounts for this drop in duplicated acquisitions, however one possible explanation is the diminished or elimination of materials budgets. In response to diminished material budgets, federal libraries may have turned their focus to processing and accessioning “hidden collections” or the receipt of gift collections.

During the period 1978-2012, no more than three percent of new cataloging records in any given year duplicated holdings held by another library. This suggests the uniqueness of each library’s collection and the value to users and the general public of having all the holdings of federal libraries available in online catalogs. It also points to the unique holdings of many of these libraries. While both the National Agricultural Library and the Forest Service Library collect materials in forestry, there is only a six percent overlap within their collections. The Forest Service Library has a large collection of information resources aggregated over the years that are unique and valuable.

Chart 4: Federal Library Duplication between 1990 - 2010 by Year



The Library of Congress benefits from the requirement (17 U.S.C. Section 407) that copyright holders/publishers must supply the Library with two copies of each published title within three months of publication. The likelihood is that a great many of the duplicate titles are held by the Library of Congress and another library.

Cooperative or shared collecting practices may be practical for the larger federal libraries, but challenging for smaller libraries with much more limited budgets and staff to process the materials. For example, the Library of Congress, National Library of Medicine and the National Agricultural Library have agreements on shared collecting policies in the areas of AIDS, biotechnology, human nutrition and veterinary science.

While acquisitions budgets are universally tight during this period of retrenchment, smaller federal libraries have fewer dollars for materials that they want to purchase. In determining what publications to buy, these libraries have to consider the expected usage of a publication, personnel costs, fees for borrowing a publication from another library, and whether borrowing from other libraries meets the timeliness requirements of its users. The costs of providing interlibrary borrowing services, e.g., processing and shipping, might not be sustainable as more libraries pursue this option. For many libraries, purchasing a duplicate copy makes sense. At

this time the volume of duplicate purchasing among federal libraries is so low that acquisitions practices and policies should be based on each library's needs.

Acquisition practices for gift materials may vary based upon the size and resources of libraries. For instance, within USDA, the library commonly receives gifts of publications from retiring employees. More recently, another means for acquisitions has been for libraries to absorb the collections of smaller agency libraries that are closing. Decisions to accept and process gifts are determined by the uniqueness of the collections, the ability to process the materials, and whether a collection needs to remain intact even if it duplicates holdings at other libraries. Retention decisions are also based on a library's mandate to provide comprehensive coverage of a subject. The study of gift acquisitions policies within federal libraries should be explored as a part of future efforts in STEM analysis.

Strategic Sourcing

FEDLINK offers collaborative acquisitions services which estimates potential savings between $127 million - $150 million annually or a savings of between nine and twenty percent for government libraries. Libraries should be encouraged to participate in joint purchasing because it may save costs for both material and acquisition functions. Some federal libraries do provide acquisitions services for their smaller agency libraries or offices, but that doesn't necessarily indicate involvement with a shared collection development strategy. It may be that there is a long standing tradition of field libraries operating independently of the main agency library.

Another area where collaboration may save money is in the area of licensed electronic resources. The complexity of this problem cannot be understated both from the standpoints of the federal libraries and the publishers. STEM titles are often the most expensive serial publications. Publishers are wary of licenses that provide access to a dispersed clientele. Many agencies have employees scattered across the country, if not across the world, which speaks to the necessity of purchasing electronic resources. A basic question that needs to be asked is whether the needs of the many federal libraries are so diverse that there is not a common corpus of publications that federal libraries could collect. If a common corpus of publications is identified, additional work is needed to identify and plan cost-effective resource sharing to the extent feasible.

There has been talk over the years of leveraging federal library serials purchasing power to lower the overall cost of online serials packages. Operationalizing this idea has been fraught with complexities. FEDLINK has begun work on this as part of its research on the Information Marketplace. A strategic sourcing initiative is underway and a STEM commodity council has been formed to bring the federal library community together to discuss challenges and opportunities. FEDLINK is facilitating and exploring collaboration with publishers of STEM materials and will continue discussions with STEM publishers and members of the federal STEM community.

Collection Management and Storage

The STEM analysis pilot of federal STEM libraries could have a number of implications in the area of collection management and digitization, specifically with regard to storage and shared

access. The implications depend upon whether or not the information gained from this type of analysis can lead federal libraries to adopt more detailed cooperative agreements. The potential exists for any cooperative agreement to affect each institution's efforts to digitize existing print resources and to make available electronic versions of their resources. The changing topical and digital landscape also implies that the process will be ongoing and iterative.

The STEM analysis allows federal libraries to know which items in their collections are duplicated at other federal libraries. The initial analysis indicates that only a relatively small amount of overlap exists between any two given collections. The exception to this is the overlap between participant institutions and the Library of Congress (see Table 2, p. 8). Even with the Library of Congress included, the overlap is relatively low, less than 30 percent for any given institution. Still, the analysis shows enough overlap to suggest some cooperation.

Strategies for cooperation among federal libraries could take multiple forms. Below, possible cooperative strategies are noted to illustrate the potential uses of this and similar data.

In the federal realm, this data could enable federal libraries to come to agreements on collective storage of and access to tangible and electronic resources. One possible strategy is for libraries to adopt cooperative agreements that delineate which topical they will be responsible for collecting in and to ensure that all participating libraries have access to those collections. Each library's policies might state that if digital copies are readily available, then there is no need to retain multiple print copies beyond what would serve as a backup. If digital copies are not available, then a certain number of copies could be retained in more than one location as long as they could be made available through interlibrary loan or electronically.

It should be noted that there are instances when duplication is desirable. For instance, many libraries provide onsite copies of dictionaries or statistical references as part of a general reference collection. There are preservation reasons for duplication as well. A variation of the shared storage strategy might be supportive of the LOCKSS (Lots of Copies Keep Stuff Safe) philosophy where multiple copies of heavily used content are collected in consideration of service issues. The higher the demand for an item, the more likely the library policies would indicate that multiple printed copies should be collected if available.

The success of an effort to consolidate and cooperate in the storage of materials would depend on the ability of participants to make publications available to each other, most likely in an electronic format. Digitizing documents into agency repositories for permanent storage and shared use could save other libraries the cost of acquiring these items.. This could eliminate duplication of paper copies among federal libraries. Having a digital archive and a digital public database would be necessary. Not all agencies have the resources to provide this type of database so a shared database that all or some federal libraries could contribute to would be necessary. The federal libraries could then also share the bibliographic records/metadata to these items, saving time and money for all. The cost to each agency could be a problem. Some type of shared agreement requiring agencies to maintain this database could be necessary.

Through cooperation, the best physical copies of any given title could be preserved while other copies could be used for digitization efforts. This would make it possible to eliminate some duplication amongst collections and strengthen each library's role as a repository of federally owned publications. In addition, some format decisions might be made based on the item itself, because it contains detailed photographs or multicolored graphic images that aren't replicated well or prohibitively expensive to digitize. If a reduction in duplicate copies is pursued, careful planning will be needed to ensure that sufficient copies exist in case a relied-upon digital copy is corrupted or lost; length or format makes an item difficult to reproduce digitally; or where usage dictates multiple copies be available.

Cooperation strategies such as these could save federal resources by reducing storage space for duplicate materials and coordinating digitization resources to focus on digitizing unique materials. While there can be some savings in cooperative arrangements, cooperation might, in the end, require more resources. Cooperation requires staff and other resources to make it happen. Each library would need to commit human resources to ensure that they meet the needs of the larger community. This might also require each agency or department to commit additional funding to pay for shared staff, equipment, and storage facilities.

Results such as those produced by this study make it possible for federal libraries to cooperate more fully and pursue multiple strategies. As the strategies suggest, more information is and will be needed moving forward. Federal libraries will need greater detail on specific items to make critical storage and digitization decisions. This pilot study only examined overlap at the title level. If cooperative work is to move ahead, federal libraries will need trusted tools and methods for recording and comparing data about such things as "best copy" or special digitization requirements, e.g., foldout maps.

Preservation

Federal libraries are facing increasing costs for housing print collections and are under pressure to reduce their footprint. At the same time, libraries are experiencing fundamental changes in user demands for new services to support the use of digital collections. These challenges as well as the opportunities to provide content through large digital collections such as HathiTrust and the U.S. Government Printing Office FDsys require libraries to rethink the management of print collections and delivery of content to their users. As federal libraries do so, they must remain cognizant of the long-term preservation of federal resources. Libraries, and federal libraries in particular, must balance the need to serve the patron of today with the obligation to offer future generations the same intellectual and creative benefits passed to this generation.

The results of the STEM collection analysis pilot have implications for a variety of issues related to preservation. Understanding the duplication across collections will assist libraries in making informed decisions about the retention and management of print copies, digitization, the provision of digital surrogates, and the preservation of digital assets.

Print Copy Management

With reliable data about the availability of print copies and digital surrogates, libraries could de-accession their duplicate copies, reducing unnecessary redundancy while ensuring that works are preserved in adequate numbers to safeguard against loss or the need for repeated digitization, and guarantee ongoing access in original form.

Data about holdings will inform discussions about collaborating on a shared print repository or network of collection spaces, allowing a coordinated, sustainable, and strategic approach to preserve federal holdings and avoid catastrophic loss of materials that could occur through an uncoordinated de-accessioning process.

As libraries assume the commitment to maintain rare collections, they will devote more resources to preserve them. With fewer copies retained, it becomes more important to follow best practices regarding environmental controls and monitoring and disaster plans. A variety of preservation actions may be applied to unique items identified and selected for retention such as digital reformatting, deacidification, stabilization, and protective housing. Shared data allows libraries to concentrate resources where they will have the greatest impact.

Better data on specific holdings and differences in condition, binding, marginalia, and other physical characteristics are required to determine which titles and how many copies need to be retained in original form.

The issue of the minimum number of copies that should be retained is complex. Libraries participating in a collaborative will need to develop policies, guidelines, and protocols for copy retention and de-accessioning, at the same time recognizing the artifactual value of specific items. Libraries will use risk management approaches to determine an acceptable level of loss.

Digital Preservation

Libraries relying on digital surrogates to take the place of print copies that have been de-accessioned or moved to off-site storage are responsible for ensuring continued access to and preservation of those digital resources. Federal libraries should explore all available avenues for collaborating to enhance the ability to preserve digital content for the long term. Libraries can become a member of National Digital Stewardship Alliance, established to maintain and advance the capacity to preserve our nation’s digital resources; participate in programs such as LOCKSS; deposit holdings in HathiTrust and FDsys; and keep abreast of new initiatives like the Digital Public Library of America.

Best Copy

If participating libraries wish to pursue collaboration on print copy management and sharing access, they will need to develop policies and procedures to determine the best copy (complete, in original format, in the best condition) among duplicates. This copy may be moved to off-site storage (dark archive), and another copy/copies could be retained for use/circulation.

It is particularly important to retain a complete copy in the best possible condition 1) if non-textual material is poorly represented in digital form; 2) to fix scanning errors, 3) as insurance against insufficient reliability of the digital copy provider.

The results of this study and future studies of this nature will have tremendous implications for federal library resource usage. The overall findings demonstrate minimum collection overlap among the participant libraries. This likely reflects either the nature of federal libraries or simply the nature of the participating libraries. Most federal libraries build collections to meet specific, specialized research needs. The relative uniqueness of the collections suggests judicious use of available resources to purchase collections.

Resources

The full resource implications are not clear at this point. However, as the community becomes more aware of the high degree of uniqueness within the federal library system, federal libraries might feel greater pressure to preserve and maintain rare collections. This could prove a resource drain. On the other hand, there might be opportunities to share storage and coordinate acquisitions. These efforts could potentially yield savings for federal libraries.

Federal library collections are an important American asset that is often overlooked by policymakers and the American public. The STEM pilot analysis brings to light the issues that ultimately impact the federal researcher, the federal scientist whose work demands access to extensive research support.

A National Asset

Information has become a commodity in the last few decades and the holdings in U.S. federal libraries represent a significant national resource. The nation's network of federal libraries manages and serves this resource on behalf of the American people. The full value and extent of this collective national asset has never been assessed. This is primarily because the federal library holdings information that is accessible tends to be only locally available and a centralized repository of federal library holdings data does not exist.

While this project has been able to compare an extensive bibliographic dataset, vast repositories of federal materials are hidden by a lack of accessible information about them. A significant portion, if not a majority, of the holdings information for participant libraries and federal libraries in general, resides in inaccessible systems or remains in a print-based format. For instance, the OCLC, which maintains the world's largest database of data about the world's libraries, has incomplete data on federal libraries as many libraries do not indicate their full holdings in OCLC systems.

At the individual collection level, a given federal library collection might not appear significant on its own, especially smaller collections. However, when combined with other federal library holding and viewed as a U.S. government asset, these largely unique federal collections comprise the richest collection of intellectual and creative output ever assembled by humankind. This project only hints at this immense American resource; its full extent has yet to be revealed.

Recommendations

The collaborators of this report reviewed the processes, data, and initial findings of this pilot, and have a number of recommendations to propose to the STEM library community. The following recommendations will be enhanced when this report and the data is made available to the library community. The utility of this information to the community is not fully understood, and this report acts as an impetus to initiate and encourage discussion about this big data issue and implications for collections management within the federal library community. The initial recommendations include:

Recommendation #1: The federal library community should pursue a comprehensive comparison of federal library holdings and develop a federal library agenda around the results of that comparison.

A comprehensive comparison of federal library holdings has never been conducted. This modest pilot suggests the value of such a comparison. A comparison can inform all aspects of federal library management and would provide the U.S. government with a richer understanding of its combined information resources.

Task 1: Dedicate long-term resources to the project.

Expanding research to include as many federal institutions as possible will require staff resources, likely two to three staff. It will also require infrastructure support including server space. The federal library community will need to complement this dedicated support with ongoing participation in projects related to future analysis projects.

Task 2: Identify tools to facilitate the efficient exchange of federal holdings information.

The ability to analyze across federal collections will be limited by the institution's ability to export records from the library catalog. In some organizations, there are insufficient resources, largely in terms of technical expertise and time, to participate in a project such as this. The holdings in library catalog include sensitive information not exportable without a "need to know". This limits the ability to conduct a comprehensive comparison across all federal library collections. If this pilot is to be expanded or built upon, technical support and related resources will need to be available.

Recommendation #2: Libraries should have their collections cataloged and inventoried so their holdings can be compared to facilitate greater cooperation with other federal libraries.

Typically, a library's holdings are entered into a large shared bibliographic utility so they are available to a wide and diverse audience of users. Libraries that have the capability should display their serial holdings in their online catalogs.

Task 1: Identify materials in the collection that are uncataloged.

This may require conducting an inventory by comparing the library's existing catalog against what's actually in the collection.

Task 2: Catalog materials according to established guidelines (e.g. AACR, RDA, Dublin Core, MARC, XML) so that records are portable and can be used to evaluate the collection against other federal libraries.

Libraries could have the option of providing full cataloging records, or minimal-level records, depending on cataloging knowledge and available resources.

Task 3: Develop ways to capture more accurate, complete holdings data.

This may require conducting an inventory by comparing the library's existing catalog against what's actually in the collection.

Recommendation #3: The federal library community should coordinate with federal agency and department leadership to ensure continued access to agency content.

Agencies are committed to ensuring their own content is available and accessible, but sometimes decision makers are unaware of how their decisions limit access. Decisions made about library collections impact the agency itself as well as any orchestrated efforts within the federal library community. All parties are committed to ensuring permanent public access to content published by the agencies.

Task 1: Increase federal awareness of library resources within and among federal institutions.

Library managers should share strategies and ideas for educating institutional management and others policy makers.

Task 2: Make transparent and collaborative decisions when those decisions have significant impact on access to agency content.

Agencies should know about and rely on the federal library community's concerted efforts to maintain access to their own agency's content. Ensuring long-term access to federal documents requires a cooperative, collaborative relationship that must be maintained.

Task 3: Share collection policies and plans with the broader federal community.

Sharing library policies, strategies, and digitization plans with your colleagues opens the door to better leveraging limited resources and eliminating redundant efforts. This can be accomplished by adding links to your library's Federal Library Directory profile or posting the documents to the FEDLIB discussion list. Collaboration between libraries that collect the same content can open doors to savings in new acquisitions and resource sharing.

Task 4: Share plans, strategies, and standards to increase access to agency funded scientific research results

The White House Office of Science and Technology Policy (OSTP) directed agencies with over $100 million in research and development (R&D) expenditures to develop draft plans for increasing access to the results those efforts by August 2013. Many agencies are looking at existing publications and/or data repositories as integral components of their plans. Sharing policies, strategies, standards, and approaches of federal libraries responding to and implementing agency plans increases the likelihood of leveraging existing efforts and providing a consistent experiences for funded researchers who will have to deposit their works, both of which are stated goals in the OSTP memo. This can be accomplished by adding links to your library's Federal Library Directory profile, creating a shared workspace in iCohere or other similar tool, or posting the documents to the FEDLIB discussion list.

Recommendation #4: Federal libraries should use a better understanding of holdings information to coordinate digitization efforts.

The federal library community has limited resources for large-scale digitization projects. The potential benefits of such projects are, however, likely to be great. This pilot has found a high degree of unique resources held by participating institutions. Many of these resources are not accessible to other federal entities, researchers and the public. Coordination among federal

libraries could help focus limited resources on efforts to digitize unique materials that might not otherwise be digitized.

Task 1: Assess the current state of federal digitization.

Federal digitization projects have largely been undertaken independently. The full extent of digitization remains unclear and would be a necessary starting point for future discussions.

Task 2: Develop a federal digitization plan and cooperative agreement that builds on the strength of each federal library.

A long-term plan and agreement will delineate responsibilities and better focus digitization resources. Moreover, a plan and agreement can serve as a valuable education tool for those unfamiliar with the vast resources held by the nation's federal libraries.

Task 3: Develop mechanisms to store and share detailed data about digitization activities.

Ongoing success of a coordinated digitization effort will be contingent upon the ability of federal libraries to access and share detailed information about specific items that have or will be digitized. No such centralized repository of this information exists for federal information. Existing tools, such as those hosted by the HathiTrust or the Government Printing Office, could be leveraged. Depending on the ultimate need determined in the planning stages, a specific federal repository may be needed.

Task 4: Use FEDLINK as a vehicle for streamlining federal digitization processes and reducing overall costs.

By building on existing initiatives such as federal scanning with Internet Archive, FEDLINK could establish contracting vehicles for participation in digital repository cooperatives, programs, and initiatives such as HathiTrust and Digital Public Library of America.

Recommendation #5: Expand research to inform a long-term preservation strategy for federal resources.

FEDLINK leads a federal library preservation group which has developed and continues to develop preservation plans and strategies. This work should be informed by the work of this pilot and related studies.

Task 1: Research the optimal number copies needed to ensure the long-term availability of federal collections.

A key issue for preservation planning is understanding the minimum number of copies that must be stored to safeguard resources for future generations. This number might vary by format and content. Many STEM related materials, for instance, have content tied to special formats, e.g., the use of foldouts in printed materials. A better understanding of the risks to this content is needed.

Task 2: Develop a clearer understanding of user behavior and needs.

Some research has been done in the area of user behavior concerning print materials and digital surrogates, but more research may be needed. Answers to key questions surrounding user behavior will inform future preservation plans. For instance, how much demand for print is there in light of digital availability? Does the presence of a digital copy increase or decrease the use of the print copy? If users appear willing to accept a digital copy, even if it is imperfect, then preservation strategies can be adapted to reflect this user behavior. Conversely, if the print is needed and expected, then strategies might need to include retention of a higher number of use copies.

Recommendation #6: Explore analyses that would benefit specific groups of libraries.

The type of collection analysis done at a 30,000 foot level has vast implications for the federal government. No acquisition decisions are currently being made at that level. Analyses done at an agency level (Army libraries) or at a Department level (Defense) could perhaps allow for more easily negotiable purchases when overlaps are exposed. Analyses could provide immediate actionable intelligence for a group of libraries in an agency or Department that is eagerly seeking such analyses.

Task 1: Solicit analyses at an agency or Department level to demonstrate how to best leverage these kinds of analyses.

Recommendation #7: Align the federal Library STEM Collection Analysis with other FEDLINK projects.

The use cases and recommendations from this pilot might be immediately helpful to inform other ongoing FEDLINK projects. Also, the work may impact other ongoing work, specifically the Federal Library Shared Collection study.

Task 1: Offer forums and conferences to facilitate communication.

Invite participants from ongoing projects to an open meeting to discuss the ongoing projects and opportunities for collaboration.

Task 2: Keep the federal library community informed of this and similar projects using the FEDLINK discussion list and press releases.

Intermittently share project updates on the FEDLIB discussion list to apprise all members of ongoing efforts especially for the member libraries that cannot participate in the annual expositions or other meetings.

Use Cases

Theoretical Applications of Data

There was active and lively discussion among the pilot participants about the ways their libraries might use the data from this pilot to inform their organization and its policies. The following use cases were developed by individual libraries with the intention of providing concrete examples, spurred by thoughts and discussion of the pilot project goals and findings. This is only a sampling of use cases. Other federal libraries may have documented use cases not included in this document.

Use Case #1

National Agricultural Library refines a strategy for selecting materials for digitization

Problem Statement

The National Agricultural Library (NAL) was created by the U.S. Department of Agriculture (USDA) in 1862 and designated a national library in 1962. Located in Beltsville, Maryland, adjacent to the USDA's Beltsville Agricultural Research Center, NAL is home to one of the world's largest collections devoted to agriculture and related sciences. With over 2.4 million volumes of books and periodicals and over 3.6 million government documents housed in its seventeen-story building, NAL also contains an ever-growing full-text digital collections and journal subscriptions. Since the library has a dual role to serve the public and the USDA, it is committed to expanding online access to its collections. Many of the library's resources and treasures can also be accessed through the NAL Digital Collections.

The NAL collection contains many unique print and serial items. In 2007, Constance Malpas of OCLC Research conducted a study of holdings by Association of Research Libraries members in OCLC's WorldCat. The results were presented by Jim Michalko to the ARL Special Collections Working Group in October 2007. (http://www.slideshare.net/oclcr/arl-uniquely-held-print-books-scwg). Malpas found that of the 125 member institutions, NAL had the 11th most unique collection. She further found that the other "national" libraries had significant unique holdings (e.g.: LC -1st and NLM- 9th). Internal sampling has also found that NAL has an additional 20,000 serial titles that are not in WorldCat and therefore may not be held anywhere else in the world. NAL also has an extensive collection of items produced by and for the Department of Agriculture. Many of these items are unique and some were not intended for wide distribution and they are in fragile condition.

Our challenge is to make NAL's print collection more accessible by digitizing it. The print collection can only be accessed by users coming to NAL. In many cases, the paper itself is deteriorating and imperiling the existence of the items themselves. These are not mutually exclusive goals. Through digitization, we can make the items widely accessible. The paper copies can then be preserved for future users.

To address these problems, NAL has committed to digitizing a sizeable portion of the print collection in the coming decades. We began with some of the popular titles of the USDA such as the Yearbook of Agriculture" and our watercolor pomology collection. The digitization was performed both in-house and by external contractors. Beginning in 2013, NAL is partnering with the Internet Archive to digitize materials at the library. The digitized items will be available both through Internet Archive and stored in our NAL Digital Collections.

Selecting which items to digitize and prioritizing the order of scanning are crucial decisions for an effective project. NAL does not want to duplicate quality scanning done by other entities. Neither does it want to miss an opportunity to scan when it has the only

available complete set of a serial title. The problem we need to address is how to determine what is unique in our collection and what others have already scanned.

Use of STEM Overlap Data

If the problem is how to determine what is duplicated in other collections and what has already been scanned, how can the STEM overlap data help?

The STEM overlap data collection aggregates the collections of eleven federal libraries including the Library of Congress, the National Library of Medicine, and the U.S. Geological Survey. The duplication among the various collections is easily identified. The data set includes the NAL bibliographic identification number, so we can precisely compare our collection with others.

The overlap data contains information on the HathiTrust holdings and that is even more important. HathiTrust combines the digitized collections from 80 different research libraries in the United States. Collectively, 68,451titles owned by NAL have been digitized by HathiTrust members. This represents 6.8% of NAL's collection. The overlap data also contains information about the provenance of the HathiTrust items. This is particularly important in determining whether the items are free of copyright restrictions.

NAL is currently working to design a workflow to integrate this information into our digitization decisions. The options include adding information on duplication to our existing catalog records. Alternatively, we can build a separate catalog that captures all the NAL items duplicated in HathiTrust and the other federal libraries. No matter which tool is chosen, the catalog will be consulted when making digitization decisions. The fact that an item is not duplicated is important as we have a high probability that no digitized copy exists. If HathiTrust has a copy, we can quickly consult the HathiTrust catalog and determine whether the digitized copy is available and complete.

Intended Outcomes

Using the data assembled by the STEM library analysis study will enable the National Agricultural Library to make informed decisions on which items to digitize. The presence of HathiTrust metadata enabled us to identify NAL items that have already been digitized. We can then avoid duplicating digitization effort, if a complete public accessible digital version held at one of the 80 HathiTrust institutions.

This type of due diligence is critical in any mass digitization project. In this case, significant time savings will be realized by having the overlap data available before we begin.

Use Case #2

NOAA makes informed collection development storage decisions

Problem Statement

Collections. The NOAA Library and Information Network (NLIN) consists of over 30 libraries scattered throughout the United States. Participation in the Network is voluntary and libraries have chosen to participate in varying degrees with the other NOAA libraries. Most, but not all, of these libraries participate in the NOAA Library and Information Network Catalog (NOAALINC). Some use NOAALINC as a complete system for their catalog, circulation, and other library functions; some use it only as a catalog; some do not use it at all. Several have their own catalog in addition to NOAALINC. As with other federal libraries, most libraries who do participate have a significant amount of uncataloged material not reflected in the NOAALINC.

The NOAA Central Library needs to make decisions about the retention and storage of its own materials, while keeping in mind the broader NOAA library community as a whole. The NOAA Central Library is located in Silver Spring, Maryland, an expensive real estate location. The Library will undergo a significant reduction (30%) in space over the next year and must make decisions about which materials to retain on open shelves, which materials to put in compact shelving, which materials to put into storage, and which materials it can excess.

To help make these decisions, the Library will need to know which materials are held by other libraries both within and outside NOAA. It will also need to know the costs of each storage option, but for the purpose of this case study, the focus will remain on the selection process.

Use of STEM Overlap Data

NOAA libraries collections and storage decisions. In addition to STEM data, NOAA will need to examine its internal holdings to determine which items are held by multiple libraries, and look at usage information to determine which items receive the most use.

The NOAA Central Library would make its storage selection decisions based upon a series of choices which determine the relative usefulness of the materials in question. The Library will use the criteria presented in the following section.

Criteria

Using the STEM analysis, the NOAA Central Library would first look at items held by other federal libraries which would fall into one of these two categories.

- *Items produced by a federal agency other than NOAA, or Items held by other federal libraries that do not fall within NOAA's main disciplines* (using a measure such as class number or other means to determine). The Library would remove the item from its collections and rely upon interlibrary loan from other agencies to serve its users.
- I*tems held by other federal libraries that do fall within NOAA's areas of expertise and disciplines or produced by NOAA or NOAA funding*. NOAA would retain the item and recommend that other libraries excess their copies. NOAA would take the lead in providing the copy for interlibrary loan either in physical form or digitized file. NOAA would determine the most cost-effective means of retention.

Once the NOAA Central Library has determined which items it will retain in its system, the Library would work with other NOAA libraries to determine where best to maintain its copies.

- *Items held in only one location within NOAA library system.* Retain the copy in the same location.
- *Items held by more than one NOAA library.* NOAA would retain only one or two copies in the most cost-effective location that can provide it via interlibrary loan and that has the means to digitize a service copy.

Each NOAA library is free to establish its own policy on retention vs. storage for those items that are unique to its library. The NOAA Central Library would have the following guidelines to determine which items to retain on its shelves and which items to store. Other NOAA libraries may choose to follow these guidelines:

- *Items held only by NOAA Central Library with low or no usage and not NOAA-produced.* The Library will store these items in compact shelving onsite.
- *Items held only by NOAA Central Library with high usage, or NOAA-produced.* The Library will retain in its main collection and digitize as resources become available.

Intended Outcomes

The NOAA Central Library will reduce the number of items it will hold on its shelves for easy access, and excess or store those items that are not deemed useful or that are available at other libraries. The STEM analysis will make the initial determination of what the Library will retain in some way. This will reduce the number of items for which further analysis is necessary, thereby reducing the workload.

The reduction in the number of physical items held by the Library will reduce the need for expensive real estate at the headquarters location, and potentially throughout NOAA. This will focus the library's human and capital resources on a smaller, but more pertinent collection of documents. This will mesh well with the Library's longer-term goal of becoming a virtual library with embedded librarians across the system.

Other factors will come into play that will affect this selection/deselection process. The Library will need to determine the actual costs of offsite storage vs. onsite storage vs. open shelf storage. The Library will have to develop a plan for digitization and long-term access to NOAA-produced documents, both retrospectively and for the future. The reduction in space dedicated to physical items and the reduction in staff time dedicated to the maintenance of physical collections will allow the Library to reassign space and staff for other purposes.

Use Case #3

Evaluating effectiveness of selection for National Forest Service Library Collection

Problem Statement

The National Forest Service Library (NFSL) was formed in 2006 from the merger of several independent libraries within the Forest Service. As a national program, NFSL is interested in systematically collecting the materials that meet the needs of our customers. There are approximately 30,000 employees in the Forest Service. NFSL has a cataloged collection of over 350,000 items with many additional uncataloged materials. The Forest Service also has an active publishing function that generates thousands of new documents every year.

Our collections consist of materials in forestry, natural resources, entomology, ecology, water, fisheries, wildlife, forest products, forest industry, outdoor recreation, wildland fire, and related materials. It includes materials authored, funded, and published by the Forest Service, as well as many other items relevant to Forest Service interests.

Identifying what to collect and obtaining those materials with limited resources is an ongoing challenge for libraries. NFSL houses many unique, rare, and fragile materials at our three permanent locations in Fort Collins, CO (headquarters); Madison, WI (forest products); and San Juan, PR (tropical forestry). A fourth location (Delaware, OH) is in the process of being closed with the collection being relocated to the other three locations.

Our primary concern is to determine whether we are collecting the materials we need to collect and not collecting materials best collected by other libraries. In addition, we want to ensure that other federal libraries have access to our resources. Lastly, we seek to validate the uniqueness of our collections.

Use of STEM Overlap Data

Because NFSL recently transitioned to an agency-wide library, we are still running two separate online catalogs. The main catalog has approximately 275,000 records, and the Forest Products Library catalog contains approximately 72,000 records. The STEM collection analysis is especially beneficial as it examined both catalogs as separate entities so we can compare between our catalogs as well as compare against other federal agency libraries and the HathiTrust Digital Library.

All federal libraries analyzed have unique and valuable collections. However, NFSL appears to have the least amount of overlap with other STEM libraries. Some of the causes may be the amount of grey literature in our collection, the number of analytical resources, and the highly specialized nature of the collections. We are frequently the only library owning an item. The results of the STEM collection analysis reaffirms how critical it is for NFSL to continue collecting in the areas we do.

As a result of this analysis, NFSL will continue to follow our current collection development policy as specified, in order of priority, below:

1. Materials authored, funded, or published by the Forest Service since the formation of the Agency in 1905 and its precursor, the Bureau of Forestry.

2. Materials about the Forest Service.

3. Materials on forestry, natural resources, and related subjects with an emphasis on North America—the materials are primarily journal literature and technical reports. This includes literature from universities, state agencies, Canadian government (local and federal), international organizations, NGO's, and consulting firms.
4. Unusual or hard to obtain resources such as proceedings and monographs.
5. A few highly relevant books.

Intended Outcomes

We hope by utilizing the STEM Collection Analysis data we will be able to:

Communicate with Agency leadership about the unique and valuable nature of this resource.

Continue following our current collection development policy that focuses on these specialized and unique materials.

Investigate preservation and conservation measures needed for so much unique material.

Identify and prioritize materials to be included in our digital repository.

Coordinate digitization efforts with the USDA National Agricultural Library.

Appendix A: Participant Email

Name	Institution	Title	Email
Carol A. Ayer	National Forest Service Library	National Program Manager	cayer@fs.fed.us
Deborah Balsamo	EPA National Library Network	National Program Manager	Balsamo.Deborah@epamail.epa.gov
Julie Blankenburg	Forest Products Laboratory, National Forest Service Library	Supervisory Librarian	jjblanke@wisc.edu
Sally Bosken	U.S. Naval Observatory Library	Director	sally.bosken@navy.mil
Christopher Cole	National Agricultural Library	Manager, Business Development	Christopher.Cole@ars.usda.gov
Blane K. Dessy	Library of Congress	Executive Director, FEDLINK	bdes@loc.gov
Thomas Doughty	Dudley Knox Library, U.S. Naval Postgraduate School	Metadata Services Librarian	todought@nps.edu
Stanley Elswick	Library Information Services Division, National Oceanic & Atmospheric Administration	Database Librarian	Stanley.Elswick@noaa.gov
Michael Esman	National Agricultural Library	Chief Collection Development Librarian	michael.esman@ars.usda.gov
Nancy Faget	Army Research Laboratory	Librarian	nancy.g.faget.civ@mail.mil
Mike Handy	Library of Congress	Deputy Assoc. Librarian, Library Services - Programs	mhan@loc.gov
Anne Harrison	Library of Congress	Librarian/Network Program Specialist, FEDLINK	anha@loc.gov
Richard L. Huffine	U.S. Geological Survey	Former U.S. Geological Survey Libraries Program	--
Neal Kaske	Library Information Services Division, National Oceanic & Atmospheric Administration	Chief, LISD	neal.kaske@noaa.gov

Irena Kavalek	U.S. Geological Survey	Supervisory Librarian	ikavalek@usgs.gov
Rosa Liu	NIST Research Library	Manager, Research Library & Information Program	rosa.liu@nist.gov
Stephen Short	Library of Congress	Program Planning Specialist	sshort@loc.gov
Jamie Stevenson	Library of Congress	Head, FEDLINK Research	jstev@loc.gov
Amanda Wilson	National Transportation Library	Director	Amanda.Wilson@dot.gov

Appendix B: Summary Overlap Charts by Library

The following charts show overall overlap information for each participant library. Each chart contains two components: a pie chart and a bar chart. The pie chart summarizes overlap as a portion of the library's entire collection. The bar chart provides information about overlap between the library named and other participant libraries.

Chart B-1: Forest Products Laboratory, National Forest Service Library

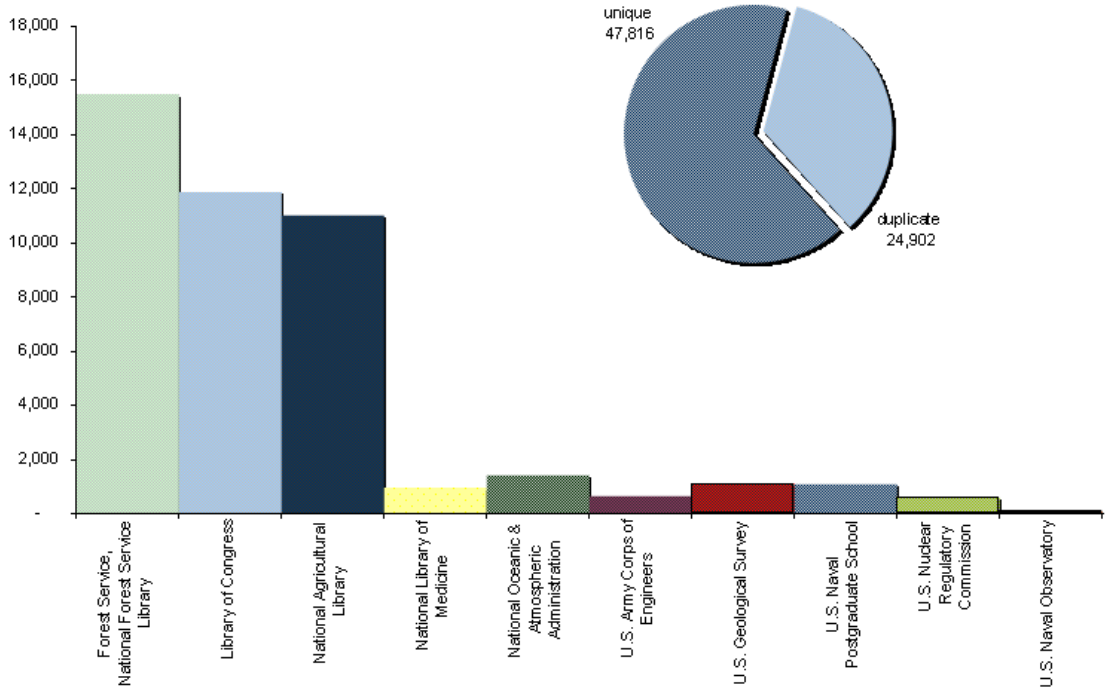


Chart B-2: Forest Service, National Forest Service Library

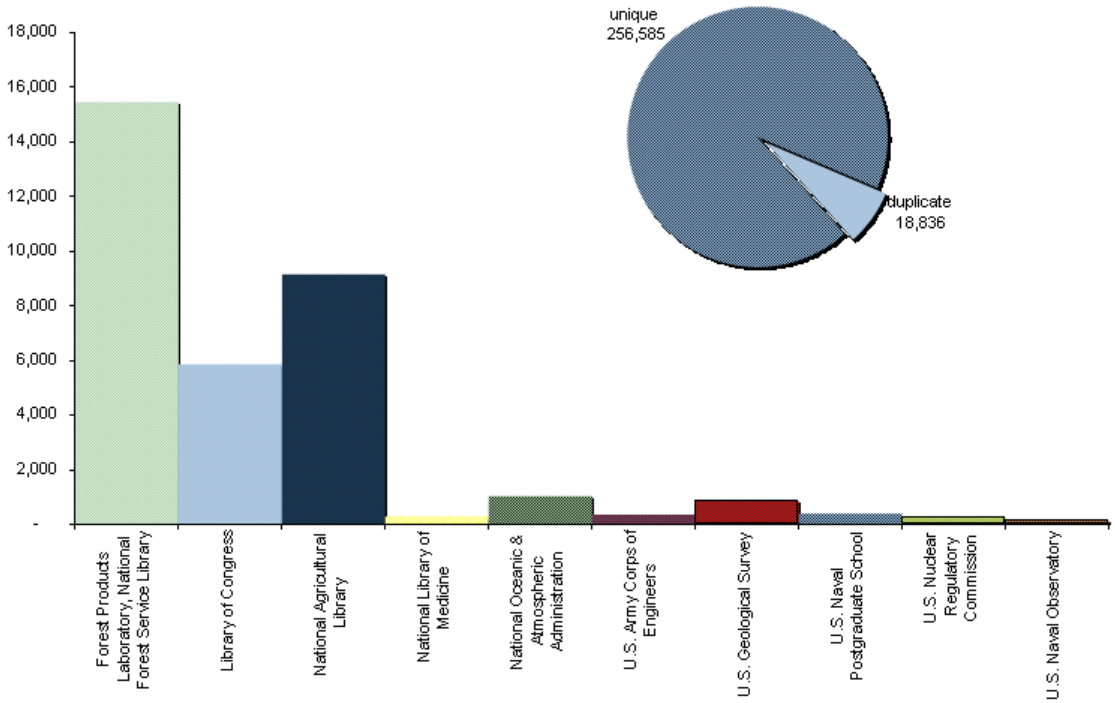


Chart B-3: Library of Congress

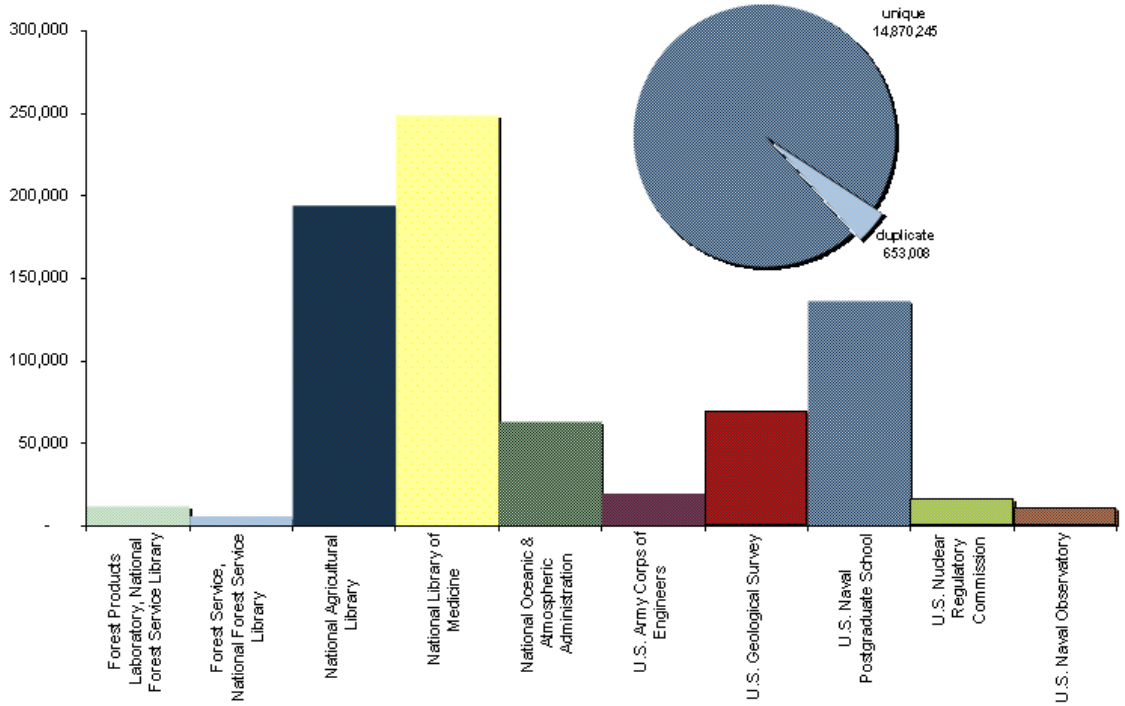


Chart B-4: National Agricultural Library

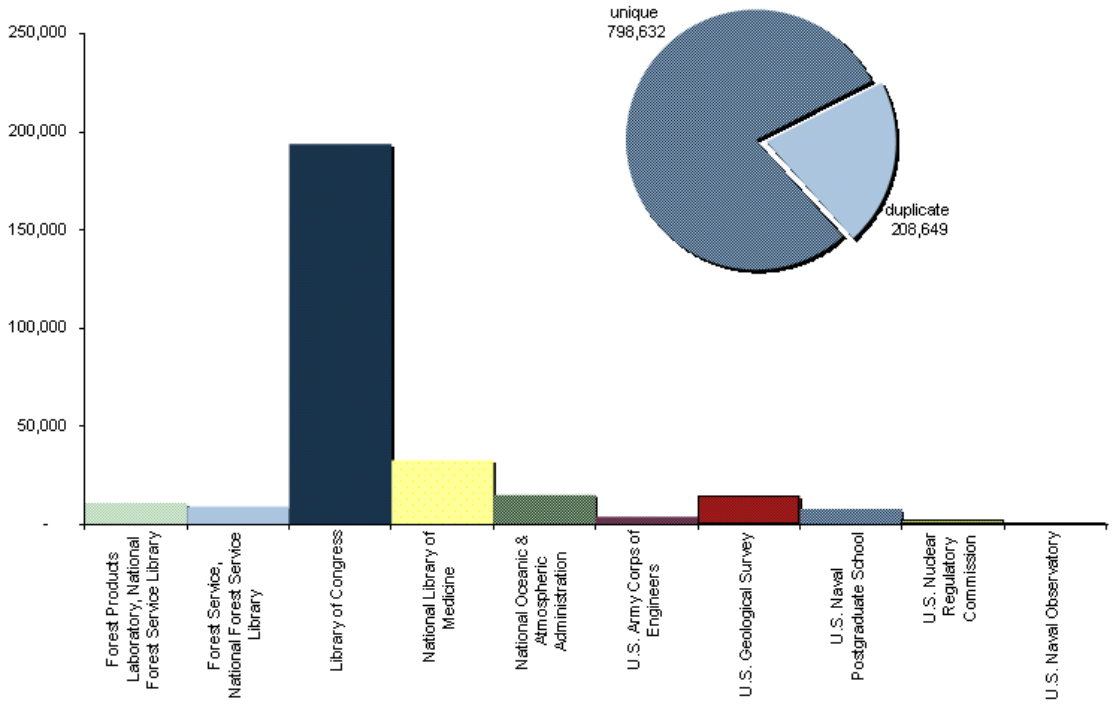


Chart B-5: National Library of Medicine

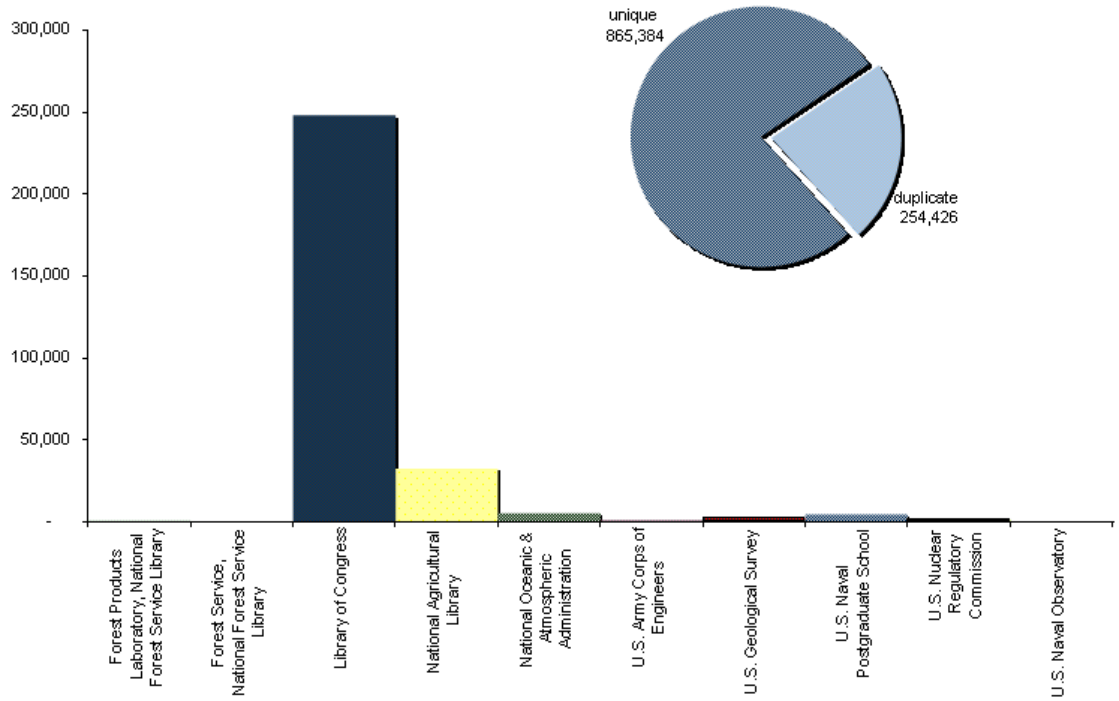


Chart B-6: National Oceanic & Atmospheric Administration National Library of Medicine

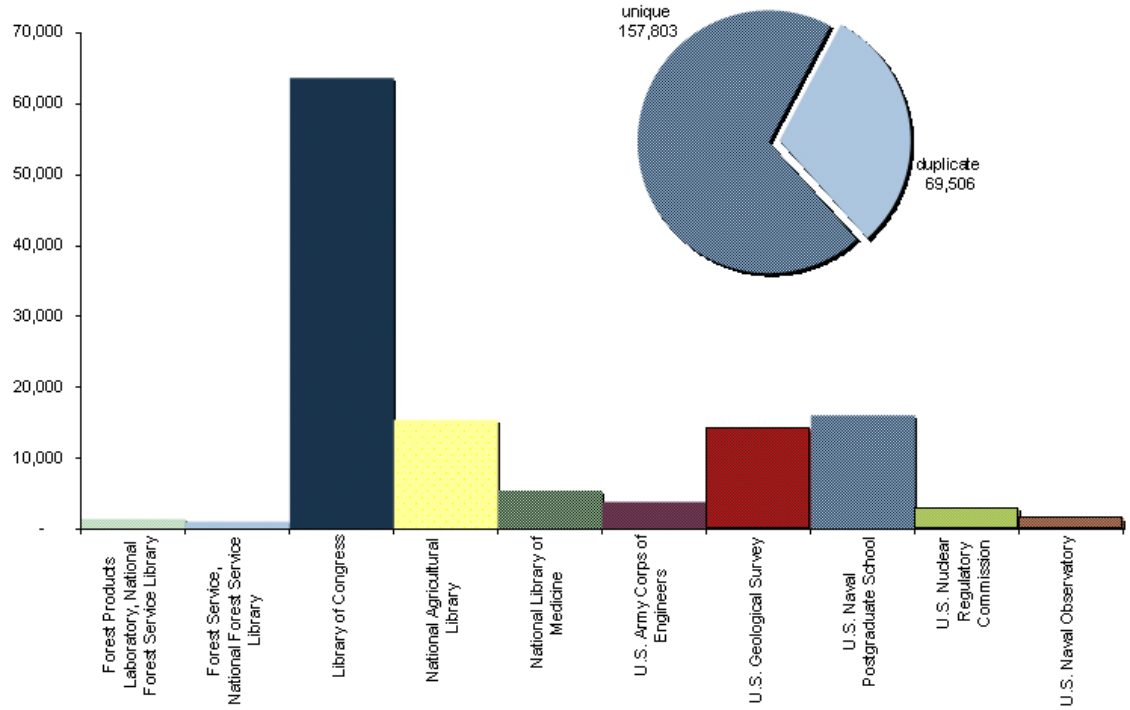


Chart B-7: U.S. Army Corps of Engineers

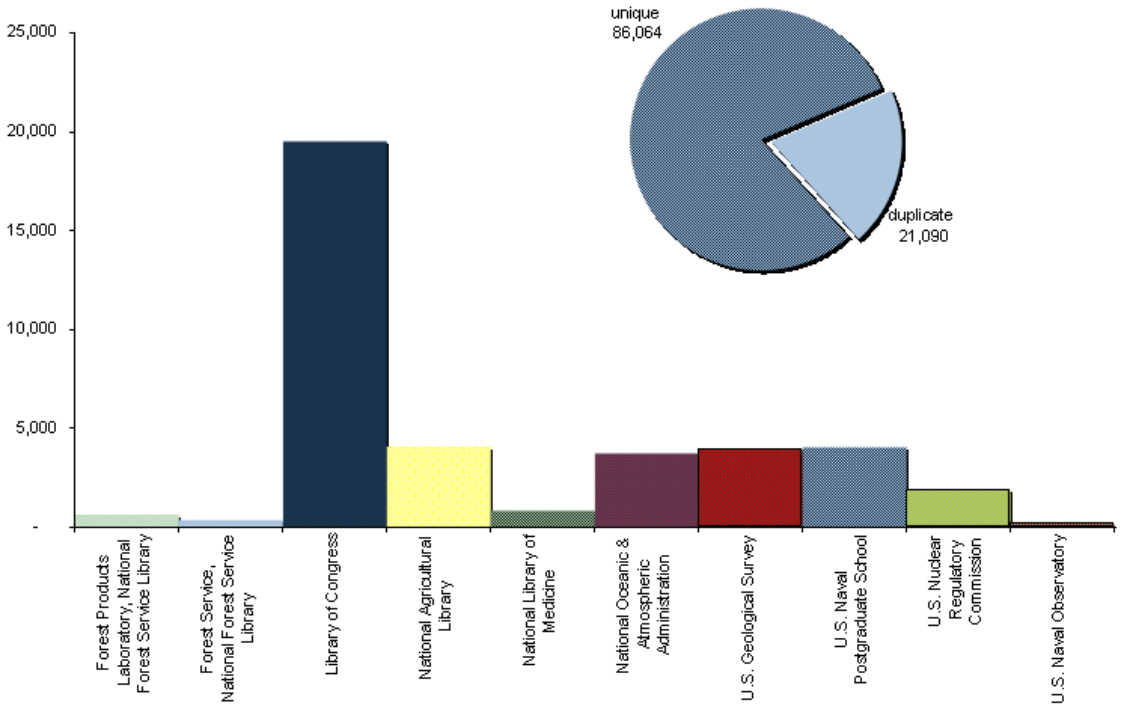


Chart B-8: U.S. Geological Survey

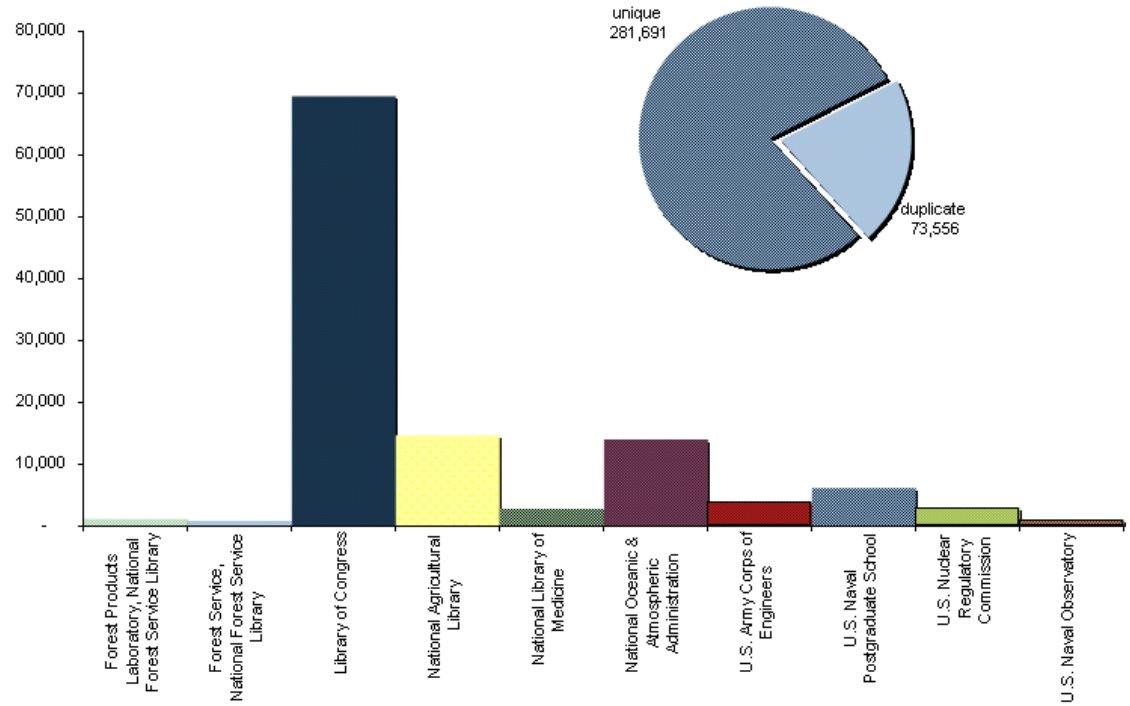


Chart B-9: U.S. Naval Postgraduate School

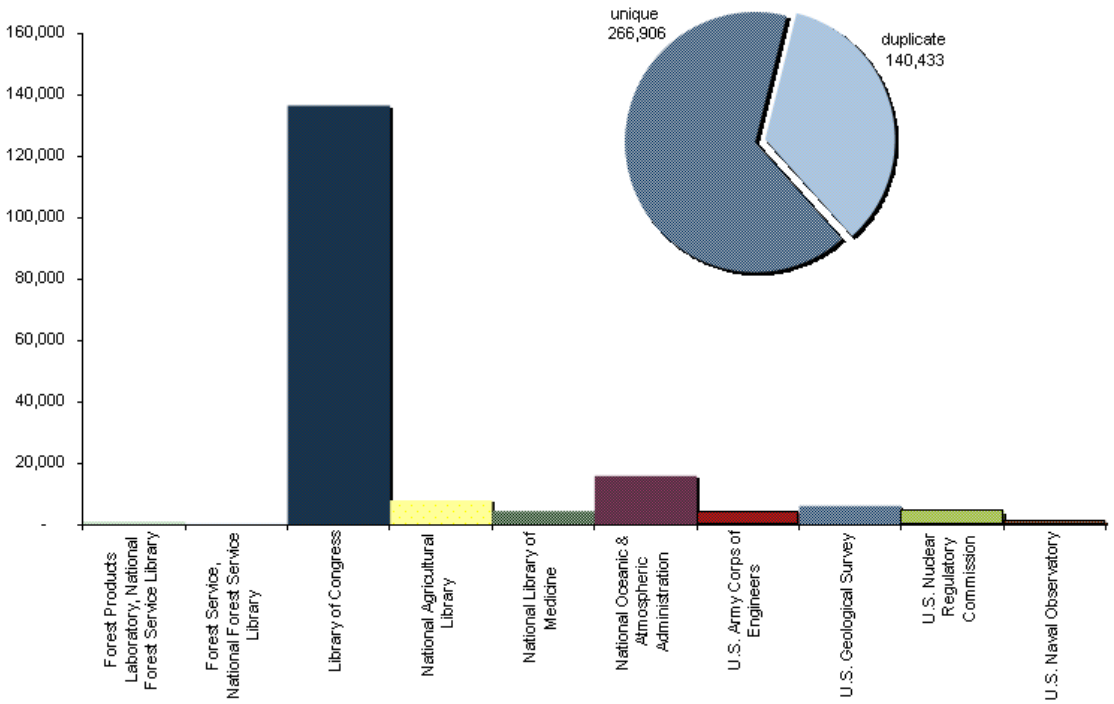


Chart B-10: U.S. Nuclear Regulatory Commission

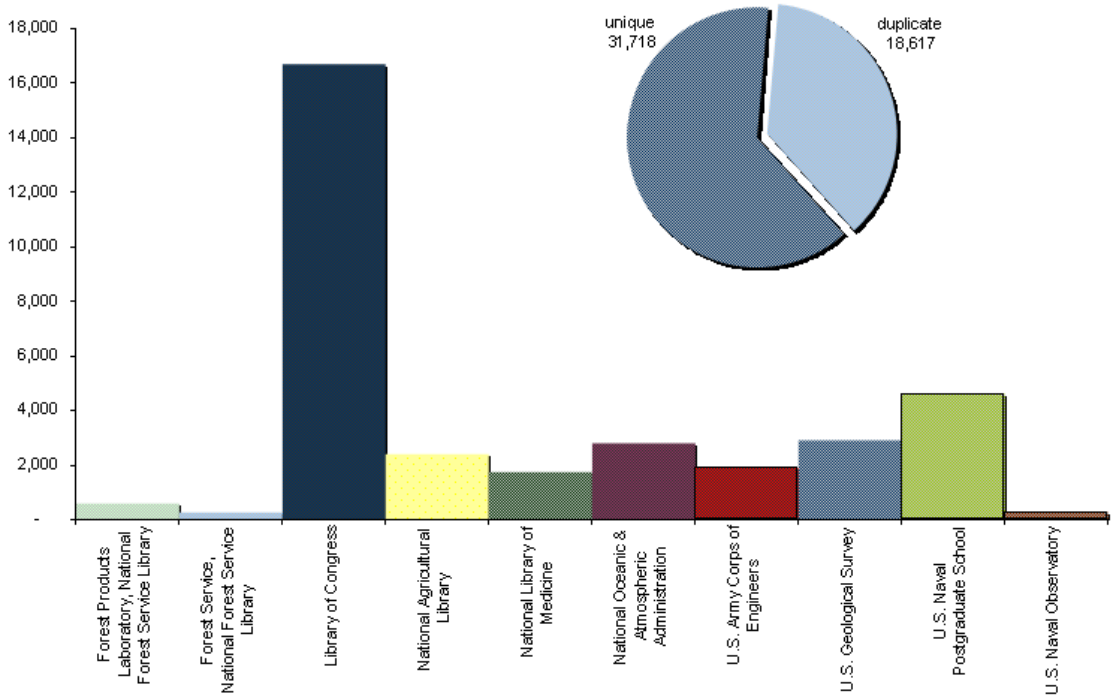
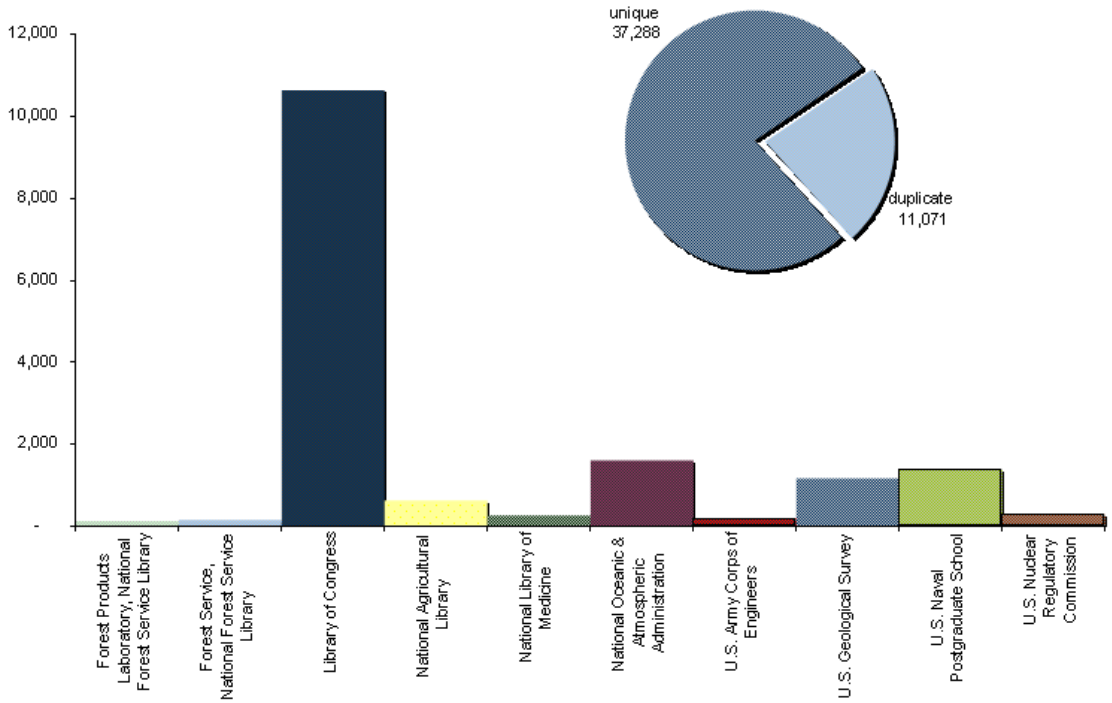


Chart B-11: U.S. Naval Observatory



Appendix C: Corresponding Data for Tables and Charts

TABLE C-1 (data for Chart 1, p. 8): Number of bibliographic records matching another library bibliographic record and HathiTrust bibliographic record

		duplication	
	total records	other pilot libraries	HathiTrust
U.S. Nuclear Regulatory Commission	50,335	18,617	15,727
U.S. Naval Postgraduate School	407,339	140,433	4,558
U.S. Naval Observatory	48,359	11,071	10,797
U.S. Geological Survey	355,247	73,556	3,387
U.S. Army Corps of Engineers	107,154	21,090	4,902
National Oceanic & Atmospheric Administration	227,309	69,506	9,675
National Library of Medicine	1,119,810	254,426	10,246
National Agricultural Library	1,007,281	208,649	200,081
Library of Congress	15,523,253	653,008	842,136
Forest Service, National Forest Service Library	275,421	18,836	20,673
Forest Products Laboratory, National Forest Service Library	72,718	24,902	6,403
All Participating Libraries	**19,194,226**	**264,767**	**39,371**

Table C-2 (data for Table 1, p. 9): Number of total bibliographic records found at other institutions

Library	Forest Products Laboratory National Forest Service Library	Forest Service, National Forest Service Library	Library of Congress	National Agricultural Library	National Library of Medicine	National Oceanic & Atmospheric Administration	U.S. Army Corps of Engineers	U.S. Geological Survey	U.S. Naval Postgraduate School	U.S. Nuclear Regulatory Commission	U.S. Naval Observatory
Forest Products Laboratory National Forest Service Library	-	19,573	11,496	11,014	980	1,411	843	1,218	1,098	597	120
Forest Service, National Forest Service Library	15,467	-	5,364	9,155	290	1,022	363	851	397	269	144
Library of Congress	13,018	5,833	-	194,272	288,963	63,347	19,628	69,705	136,385	16,850	18,736
National Agricultural Library	11,050	9,176	193,799	-	32,661	15,127	4,082	14,595	7,972	2,439	598
National Library of Medicine	962	592	248,374	31,908		5,892	842	2,782	4,494	1,799	258
National Oceanic & Atmospheric Administration	1,468	1,078	63,568	15,402	5,459	-	3,853	14,204	16,043	2,924	1,618
U.S. Army Corps of Engineers	843	565	19,917	4,073	843	3,747	-	1,941	4,051	1,805	189
U.S. Geological Survey	1,140	854	69,522	14,666	2,808	14,010	3,950	-	6,161	2,970	1,058
U.S. Naval Postgraduate School	1,098	481	136,532	8,123	4,917	15,917	4,144	6,221	-	4,637	1,307
U.S. Nuclear Regulatory Commission	594	266	16,701	2,408	1,748	2,801	1,898	2,911	4,612	-	271
U.S. Naval Observatory	171	156	18,197	627	258	1,507	192	1,161	1,309	281	

Table C-3 (data for Chart 2, p. 10): Federal library duplication by year

year	total records	Dupe -licates	year	total records	Dupe -licates	year	total records	Dupe -licates
1501	2,057,665	0	1546	2,057,665	0	1591	2,057,665	0
1502	2,057,665	0	1547	2,057,665	0	1592	2,057,665	0
1503	2,057,665	0	1548	2,057,665	0	1593	2,057,665	0
1504	2,057,665	0	1549	2,057,665	0	1594	2,057,665	0
1505	2,057,665	0	1550	2,057,665	0	1595	2,057,665	0
1506	2,057,665	0	1551	2,057,665	0	1596	2,057,665	0
1507	2,057,665	0	1552	2,057,665	0	1597	2,057,665	0
1508	2,057,665	0	1553	2,057,665	0	1598	2,057,665	0
1509	2,057,665	0	1554	2,057,665	0	1599	2,057,665	2
1510	2,057,665	0	1555	2,057,665	0	1600	2,057,665	0
1511	2,057,665	0	1556	2,057,665	0	1601	2,057,665	2
1512	2,057,665	0	1557	2,057,665	0	1602	2,057,665	0
1513	2,057,665	0	1558	2,057,665	0	1603	2,057,665	0
1514	2,057,665	0	1559	2,057,665	0	1604	2,057,665	0
1515	2,057,665	0	1560	2,057,665	0	1605	2,057,665	0
1516	2,057,665	0	1561	2,057,665	0	1606	2,057,665	0
1517	2,057,665	0	1562	2,057,665	0	1607	2,057,665	0
1518	2,057,665	1	1563	2,057,665	0	1608	2,057,665	0
1519	2,057,665	0	1564	2,057,665	0	1609	2,057,665	0
1520	2,057,665	0	1565	2,057,665	0	1610	2,057,665	0
1521	2,057,665	0	1566	2,057,665	0	1611	2,057,665	0
1522	2,057,665	0	1567	2,057,665	0	1612	2,057,665	0
1523	2,057,665	0	1568	2,057,665	0	1613	2,057,665	0
1524	2,057,665	0	1569	2,057,665	0	1614	2,057,665	0
1525	2,057,665	0	1570	2,057,665	0	1615	2,057,665	2
1526	2,057,665	0	1571	2,057,665	0	1616	2,057,665	0
1527	2,057,665	0	1572	2,057,665	2	1617	2,057,665	2
1528	2,057,665	0	1573	2,057,665	0	1618	2,057,665	0
1529	2,057,665	0	1574	2,057,665	0	1619	2,057,665	0
1530	2,057,665	0	1575	2,057,665	0	1620	2,057,665	0
1531	2,057,665	0	1576	2,057,665	0	1621	2,057,665	0
1532	2,057,665	0	1577	2,057,665	0	1622	2,057,665	0
1533	2,057,665	0	1578	2,057,665	0	1623	2,057,665	0
1534	2,057,665	0	1579	2,057,665	0	1624	2,057,665	0
1535	2,057,665	0	1580	2,057,665	0	1625	2,057,665	2
1536	2,057,665	0	1581	2,057,665	2	1626	2,057,665	0
1537	2,057,665	0	1582	2,057,665	0	1627	2,057,665	0
1538	2,057,665	0	1583	2,057,665	0	1628	2,057,665	0
1539	2,057,665	0	1584	2,057,665	2	1629	2,057,665	0
1540	2,057,665	0	1585	2,057,665	0	1630	2,057,665	0
1541	2,057,665	0	1586	2,057,665	0	1631	2,057,665	2
1542	2,057,665	0	1587	2,057,665	0	1632	2,057,665	0
1543	2,057,665	0	1588	2,057,665	0	1633	2,057,665	0
1544	2,057,665	0	1589	2,057,665	0	1634	2,057,665	0
1545	2,057,665	0	1590	2,057,665	2	1635	2,057,665	0

year	total records	Dupe -licates
1636	2,057,665	0
1637	2,057,665	0
1638	2,057,665	0
1639	2,057,665	0
1640	2,057,665	2
1641	2,057,665	0
1642	2,057,665	0
1643	2,057,665	0
1644	2,057,665	0
1645	2,057,665	0
1646	2,057,665	0
1647	2,057,665	2
1648	2,057,665	0
1649	2,057,665	0
1650	2,057,665	0
1651	2,057,665	0
1652	2,057,665	1
1653	2,057,665	0
1654	2,057,665	2
1655	2,057,665	0
1656	2,057,665	2
1657	2,057,665	0
1658	2,057,665	0
1659	2,057,665	0
1660	2,057,665	2
1661	2,057,665	0
1662	2,057,665	0
1663	2,057,665	0
1664	2,057,665	2
1665	2,057,665	3
1666	2,057,665	4
1667	2,057,665	0
1668	2,057,665	0
1669	2,057,665	0
1670	2,057,665	1
1671	2,057,665	4
1672	2,057,665	3
1673	2,057,665	2
1674	2,057,665	2
1675	2,057,665	0
1676	2,057,665	0
1677	2,057,665	0
1678	2,057,665	2
1679	2,057,665	2
1680	2,057,665	0
1681	2,057,665	0
1682	2,057,665	3
1683	2,057,665	5
1684	2,057,665	0
1685	2,057,665	4
1686	2,057,665	0
1687	2,057,665	0
1688	2,057,665	0
1689	2,057,665	2
1690	2,057,665	2
1691	2,057,665	0
1692	2,057,665	2
1693	2,057,665	2
1694	2,057,665	0
1695	2,057,665	0
1696	2,057,665	0
1697	2,057,665	2
1698	2,057,665	0
1699	2,057,665	5
1700	2,057,665	1
1701	2,057,665	2
1702	2,057,665	0
1703	2,057,665	2
1704	2,057,665	0
1705	2,057,665	6
1706	2,057,665	0
1707	2,057,665	0
1708	2,057,665	0
1709	2,057,665	4
1710	2,057,665	0
1711	2,057,665	2
1712	2,057,665	2
1713	2,057,665	5
1714	2,057,665	0
1715	2,057,665	0
1716	2,057,665	0
1717	2,057,665	6
1718	2,057,665	4
1719	2,057,665	0
1720	2,057,665	0
1721	2,057,665	2
1722	2,057,665	1
1723	2,057,665	2
1724	2,057,665	2
1725	2,057,665	4
1726	2,057,665	0
1727	2,057,665	4
1728	2,057,665	0
1729	2,057,665	4
1730	2,057,665	0
1731	2,057,665	0
1732	2,057,665	2
1733	2,057,665	2
1734	2,057,665	0
1735	2,057,665	0
1736	2,057,665	0
1737	2,057,665	3
1738	2,057,665	1
1739	2,057,665	4
1740	2,057,665	4
1741	2,057,665	4
1742	2,057,665	3
1743	2,057,665	0
1744	2,057,665	2
1745	2,057,665	11
1746	2,057,665	0
1747	2,057,665	2
1748	2,057,665	10
1749	2,057,665	6
1750	2,057,665	10
1751	2,057,665	4
1752	2,057,665	2
1753	2,057,665	4
1754	2,057,665	4
1755	2,057,665	2
1756	2,057,665	3
1757	2,057,665	6
1758	2,057,665	4
1759	2,057,665	8
1760	2,057,665	4
1761	2,057,665	0
1762	2,057,665	10
1763	2,057,665	11
1764	2,057,665	4
1765	2,057,665	0
1766	2,057,665	4
1767	2,057,665	8
1768	2,057,665	8
1769	2,057,665	4
1770	2,057,665	8
1771	2,057,665	20
1772	2,057,665	7
1773	2,057,665	12
1774	2,057,665	10
1775	2,057,665	16
1776	2,057,665	17
1777	2,057,665	8
1778	2,057,665	12
1779	2,057,665	2

year	total records	Dupe -licates
1780	2,057,665	6
1781	2,057,665	4
1782	2,057,665	0
1783	2,057,665	14
1784	2,057,665	17
1785	2,057,665	13
1786	2,057,665	6
1787	2,057,665	12
1788	2,057,665	10
1789	2,057,665	17
1790	2,057,665	0
1791	2,057,665	10
1792	2,057,665	10
1793	2,057,665	8
1794	2,057,665	28
1795	2,057,665	13
1796	2,057,665	8
1797	2,057,665	4
1798	2,057,665	13
1799	2,057,665	19
1800	2,057,665	23
1801	2,057,665	31
1802	2,057,665	16
1803	2,057,665	17
1804	2,057,665	21
1805	2,057,665	15
1806	2,057,665	13
1807	2,057,665	24
1808	2,057,665	19
1809	2,057,665	28
1810	2,057,665	20
1811	2,057,665	52
1812	2,057,665	23
1813	2,057,665	19
1814	2,057,665	50
1815	2,057,665	22
1816	2,057,665	35
1817	2,057,665	38
1818	2,057,665	58
1819	2,057,665	48
1820	2,057,665	51
1821	2,057,665	54
1822	2,057,665	57
1823	2,057,665	57
1824	2,057,665	78
1825	2,057,665	66
1826	2,057,665	67
1827	2,057,665	46
1828	2,057,665	64
1829	2,057,665	80
1830	2,057,665	70
1831	2,057,665	72
1832	2,057,665	89
1833	2,057,665	64
1834	2,057,665	94
1835	2,057,665	92
1836	2,057,665	77
1837	2,057,665	85
1838	2,057,665	76
1839	2,057,665	116
1840	2,057,665	128
1841	2,057,665	94
1842	2,057,665	115
1843	2,057,665	120
1844	2,057,665	105
1845	2,057,665	159
1846	2,057,665	159
1847	2,057,665	99
1848	2,057,665	97
1849	2,057,665	146
1850	2,057,665	158
1851	2,057,665	141
1852	2,057,665	189
1853	2,057,665	167
1854	2,057,665	167
1855	2,057,665	186
1856	2,057,665	193
1857	2,057,665	156
1858	2,057,665	152
1859	2,057,665	190
1860	2,057,665	188
1861	2,057,665	144
1862	2,057,665	148
1863	2,057,665	133
1864	2,057,665	141
1865	2,057,665	186
1866	2,057,665	255
1867	2,057,665	222
1868	2,057,665	220
1869	2,057,665	214
1870	2,057,665	230
1871	2,057,665	211
1872	2,057,665	253
1873	2,057,665	283
1874	2,057,665	268
1875	2,057,665	278
1876	2,057,665	293
1877	2,057,665	233
1878	2,057,665	289
1879	2,057,665	279
1880	2,057,665	366
1881	2,057,665	285
1882	2,057,665	381
1883	2,057,665	434
1884	2,057,665	466
1885	2,057,665	408
1886	2,057,665	397
1887	2,057,665	397
1888	2,057,665	412
1889	2,057,665	414
1890	2,057,665	437
1891	2,057,665	475
1892	2,057,665	419
1893	2,057,665	602
1894	2,057,665	512
1895	2,057,665	558
1896	2,057,665	591
1897	2,057,665	639
1898	2,057,665	605
1899	2,057,665	596
1900	2,057,665	1285
1901	2,057,665	831
1902	2,057,665	852
1903	2,057,665	881
1904	2,057,665	921
1905	2,057,665	881
1906	2,057,665	864
1907	2,057,665	931
1908	2,057,665	958
1909	2,057,665	989
1910	2,057,665	1100
1911	2,057,665	1374
1912	2,057,665	1510
1913	2,057,665	1628
1914	2,057,665	1497
1915	2,057,665	1391
1916	2,057,665	1585
1917	2,057,665	1685
1918	2,057,665	1524
1919	2,057,665	1594
1920	2,057,665	1773
1921	2,057,665	1855
1922	2,057,665	2179
1923	2,057,665	2145

year	total records	Dupe -licates
1924	2,057,665	2012
1925	2,057,665	2228
1926	2,057,665	2418
1927	2,057,665	2321
1928	2,057,665	2374
1929	2,057,665	2336
1930	2,057,665	2812
1931	2,057,665	2769
1932	2,057,665	2599
1933	2,057,665	2326
1934	2,057,665	2752
1935	2,057,665	3032
1936	2,057,665	3395
1937	2,057,665	3626
1938	2,057,665	3817
1939	2,057,665	4245
1940	2,057,665	4260
1941	2,057,665	4391
1942	2,057,665	4404
1943	2,057,665	4036
1944	2,057,665	3097
1945	2,057,665	3071
1946	2,057,665	3196
1947	2,057,665	3604
1948	2,057,665	3546
1949	2,057,665	4045
1950	2,057,665	4353
1951	2,057,665	3867
1952	2,057,665	3683
1953	2,057,665	3823
1954	2,057,665	3687
1955	2,057,665	3559
1956	2,057,665	3604
1957	2,057,665	3634
1958	2,057,665	4065
1959	2,057,665	4359
1960	2,057,665	5423
1961	2,057,665	5315
1962	2,057,665	6041
1963	2,057,665	7238
1964	2,057,665	8469
1965	2,057,665	7846
1966	2,057,665	8464
1967	2,057,665	11026
1968	2,057,665	15293
1969	2,057,665	22613
1970	2,057,665	28423
1971	2,057,665	32765
1972	2,057,665	37157
1973	2,057,665	40653
1974	2,057,665	44517
1975	2,057,665	50984
1976	2,057,665	51311
1977	2,057,665	51059
1978	2,057,665	48800
1979	2,057,665	46155
1980	2,057,665	44296
1981	2,057,665	44880
1982	2,057,665	42529
1983	2,057,665	42283
1984	2,057,665	44642
1985	2,057,665	43208
1986	2,057,665	42252
1987	2,057,665	43990
1988	2,057,665	47146
1989	2,057,665	49537
1990	2,057,665	52993
1991	2,057,665	51110
1992	2,057,665	50431
1993	2,057,665	48128
1994	2,057,665	49316
1995	2,057,665	47918
1996	2,057,665	47797
1997	2,057,665	45140
1998	2,057,665	43449
1999	2,057,665	43243
2000	2,057,665	41937
2001	2,057,665	40944
2002	2,057,665	40249
2003	2,057,665	39778
2004	2,057,665	38626
2005	2,057,665	39158
2006	2,057,665	43308
2007	2,057,665	39067
2008	2,057,665	33118
2009	2,057,665	29217
2010	2,057,665	23765

Table C-4 (data for Chart 3, p. 11): Time distribution of federal record matches with HathiTrust records

Year	# of duplicates
1502	1
1504	1
1509	1
1521	1
1528	1
1530	1
1537	2
1538	2
1539	2
1541	1
1542	3
1543	2
1545	2
1549	1
1550	1
1551	1
1553	1
1554	2
1555	1
1557	1
1558	1
1559	1
1560	2
1561	1
1564	2
1566	1
1567	2
1568	1
1571	1
1572	2
1573	2
1576	1
1577	1
1578	1
1582	1
1583	1
1585	1
1586	1
1591	2
1593	1
1596	1
1599	4
1600	2
1601	1
1605	2
1608	1
1610	5
1612	1
1613	1
1618	1
1619	1
1621	1
1623	5
1625	3
1626	2
1628	2
1631	2
1634	1
1636	1
1638	1
1639	2
1640	3
1641	2
1643	3
1644	1
1645	3
1646	1
1647	1
1648	1
1650	2
1651	2
1652	3
1653	1
1654	2
1656	3
1657	2
1658	2
1660	6
1662	2
1663	2
1664	3
1665	9
1666	1
1667	2
1669	7
1670	3
1671	1
1672	5
1673	5
1674	2
1675	5
1676	4
1677	2
1678	1
1679	2
1680	4
1681	4
1682	5
1683	3
1684	4
1685	5
1686	3
1687	6
1688	5
1689	3
1690	7
1691	6
1692	2
1693	1
1694	10
1695	8
1696	4
1697	3
1698	5
1699	6
1700	14
1701	10
1702	12
1703	4
1704	3
1705	4
1706	3
1707	4
1708	6
1709	9
1710	9
1711	7
1712	8
1713	5
1714	14
1715	6
1716	7
1717	9
1718	6
1719	6
1720	16
1721	9
1722	11
1723	8
1724	15
1725	10
1726	14
1727	17
1728	10
1729	14
1730	14
1731	10
1732	18
1733	9
1734	12
1735	17
1736	6
1737	10
1738	13
1739	19
1740	22
1741	13
1742	14
1743	15
1744	18
1745	14
1746	9
1747	16
1748	13
1749	19
1750	26
1751	19
1752	23
1753	28
1754	34
1755	29
1756	17
1757	18
1758	23
1759	36
1760	28
1761	23
1762	30
1763	30
1764	13

Year	# of duplicates
1765	10
1766	21
1767	29
1768	34
1769	29
1770	44
1771	43
1772	30
1773	26
1774	29
1775	32
1776	34
1777	46
1778	40
1779	34
1780	37
1781	30
1782	20
1783	47
1784	37
1785	34
1786	34
1787	50
1788	63
1789	61
1790	35
1791	64
1792	55
1793	64
1794	84
1795	49
1796	44
1797	56
1798	48
1799	60
1800	306
1801	127
1802	127
1803	106
1804	132
1805	119
1806	149
1807	153
1808	153
1809	144
1810	170
1811	152
1812	151
1813	160
1814	167
1815	166
1816	182
1817	191
1818	234
1819	216
1820	279
1821	257
1822	319
1823	307
1824	330
1825	355
1826	332
1827	307
1828	366
1829	415
1830	396
1831	333
1832	387
1833	431
1834	459
1835	484
1836	510
1837	423
1838	496
1839	561
1840	603
1841	528
1842	475
1843	541
1844	595
1845	709
1846	693
1847	706
1848	676
1849	627
1850	839
1851	816
1852	865
1853	878
1854	904
1855	890
1856	1,033
1857	877
1858	831
1859	929
1860	1,101
1861	906
1862	880
1863	969
1864	1,039
1865	1,192
1866	1,067
1867	1,148
1868	1,026
1869	1,033
1870	1,033
1871	999
1872	966
1873	955
1874	1,056
1875	1,052
1876	1,210
1877	1,160
1878	1,151
1879	1,265
1880	1,387
1881	1,332
1882	1,392
1883	1,531
1884	1,542
1885	1,462
1886	1,576
1887	1,760
1888	1,756
1889	1,806
1890	2,012
1891	1,867
1892	2,033
1893	2,165
1894	2,063
1895	2,459
1896	2,463
1897	2,536
1898	2,685
1899	2,877
1900	3,963
1901	3,281
1902	3,852
1903	3,916
1904	3,923
1905	3,923
1906	4,050
1907	4,411
1908	4,406
1909	4,566
1910	5,107
1911	5,386
1912	5,601
1913	5,700
1914	5,562
1915	5,489
1916	5,503
1917	5,579
1918	5,295
1919	5,473
1920	6,367
1921	5,847
1922	6,712
1923	5,582
1924	5,327
1925	5,724
1926	5,786
1927	6,259
1928	6,426
1929	6,482
1930	7,353
1931	6,844
1932	6,147
1933	6,099
1934	6,677
1935	7,107
1936	7,364
1937	7,473
1938	7,593
1939	7,704
1940	7,997
1941	7,252
1942	7,145
1943	7,383
1944	6,839
1945	7,264
1946	8,724
1947	9,232
1948	9,447
1949	9,863
1950	10,316
1951	9,528
1952	9,474
1953	9,664
1954	9,904
1955	10,316
1956	10,093

Year	# of duplicates
1957	10,347
1958	10,786
1959	11,789
1960	14,464
1961	14,640
1962	17,338
1963	18,640
1964	19,371
1965	21,315
1966	22,173
1967	19,770
1968	17,382
1969	17,142
1970	17,534
1971	18,030
1972	16,493
1973	13,761
1974	10,716
1975	11,028
1976	9,584
1977	9,074
1978	8,742
1979	7,509
1980	9,372
1981	7,634
1982	6,180
1983	6,272
1984	6,014
1985	4,580
1986	4,378
1987	5,352
1988	6,152
1989	6,924
1990	7,635
1991	6,717
1992	7,644
1993	10,591
1994	11,274
1995	10,888
1996	10,839
1997	13,231
1998	13,264
1999	11,807
2000	10,775
2001	9,555
2002	10,177
2003	11,907
2004	13,399

Year	# of duplicates
2005	14,458
2006	18,107
2007	12,411
2008	7,741
2009	3,592
2010	1,531

Table C-5 (data for Chart 4, p. 13): Federal library duplication by year, 1990-2010

year	total records	duplicates
1990	301,633	52,993
1991	298,519	51,110
1992	296,911	50,431
1993	297,452	48,128
1994	297,052	49,316
1995	289,512	47,918
1996	290,183	47,797
1997	296,513	45,140
1998	309,472	43,449
1999	312,737	43,243
2000	327,825	41,937
2001	313,719	40,944
2002	319,832	40,249
2003	311,438	39,778
2004	314,273	38,626
2005	318,959	39,158
2006	318,932	43,308
2007	320,511	39,067
2008	307,911	33,118
2009	303,601	29,217
2010	287,593	23,765