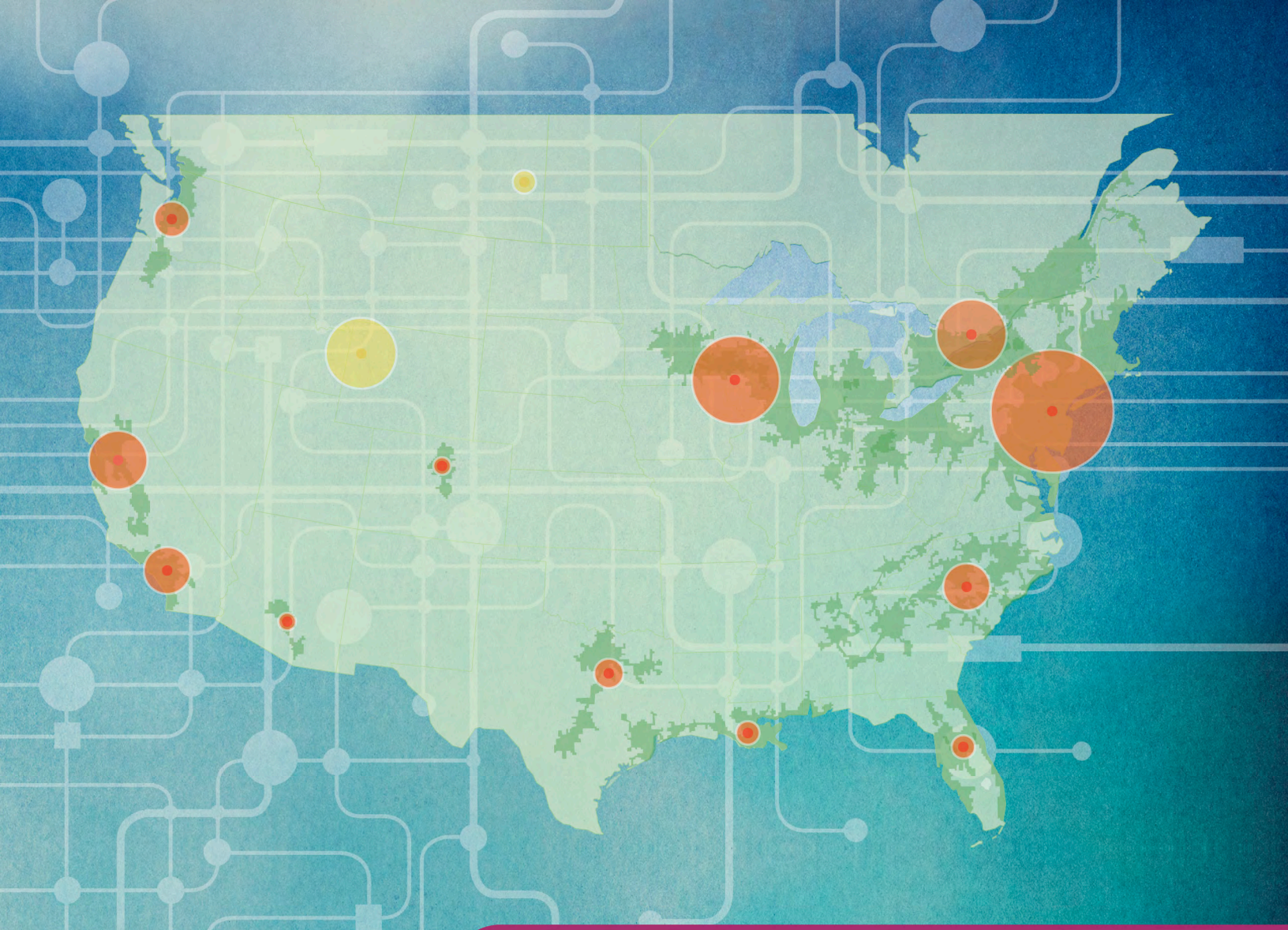


Understanding the Collective Collection:

Towards a System-wide Perspective on Library Print Collections

By Lorcan Dempsey, Brian Lavoie and Constance Malpas
with Lynn Silipigni Connaway, Roger C. Schonfeld, JD Shipengrover and Günter Waibel

OCLC Research

Understanding the Collective Collection: Towards a System-wide Perspective on Library Print Collections

By Lorcan Dempsey, Brian Lavoie and Constance Malpas

with Lynn Silipigni Connaway, Roger C. Schonfeld, JD Shipengrover and Günter Waibel

December 2013

OCLC Research
Dublin, Ohio 43017 USA
www.oclc.org

ISBN: 1-55653-461-2 (978-1-55653-461-4)
DOI: 10.25333/E94Q-9Q39
OCLC (WorldCat): 855271983

Please direct correspondence to:
OCLC Research
oclcresearch@oclc.org

Suggested citation:
Dempsey, Lorcan, Brian Lavoie, Constance Malpas, Lynn Silipigni Connaway, Roger C. Schonfeld, JD Shipengrover, and Günter Waibel. 2013. *Understanding the Collective Collection: Towards a System-wide Perspective on Library Print Collections*. Dublin, Ohio: OCLC Research. https://doi.org/10.25333/E94Q-9Q39.

Contents

Acknowledgments

Special thanks to the following OCLC staff for their help with compiling and publishing this report: Melissa Renspie, Jeanette McNicol, Eric Childress, Karen Disbrow, Reneé Page and Brad Gauder.

Introduction

The Emergence of the Collective Collection:
Analyzing Aggregate Print Library Holdings

Lorcan Dempsey

As the network continues to reconfigure personal, business and institutional relationships, it is natural that we also continue to see changes in how library collections are managed: changes in focus, boundaries and value.

One important trend is that libraries and the organizations that provide services to them will devote more attention to system-wide organization of collections—whether the "system" is a consortium, a region or a country.

We have become used to this in the digital environment, where the scale advantage of such consolidation is apparent. Think of shared approaches to preservation of licensed material, as in CLOCKSS or Portico, for example. Or think of the emergence of JSTOR and HathiTrust, under very different business models, as shared digital hubs that concentrate capacity—a natural trend as materials are digitized and aggregated at the network level. Or think of the interest in aggregating metadata across institutional digital collections, as in Europeana,[1] WorldCat or the Digital Public Library of America.[2]

Recently, print collections have also been the subject of such shared attention. Libraries are beginning to evolve arrangements that will facilitate long-term shared management of the print literature as individual libraries begin to manage down their local capacity. Examples of initiatives here are the WEST Project[3] and the print management activities of the HathiTrust. Initially, attention was focused on journal runs, but it is now spreading to monographs, as well. Of course, libraries have long worked with print repositories, individually or in shared settings. However, a more systemic perspective is now emerging and we have been using the phrase "collective collection" to evoke this more focused attention on collective development, management and disclosure of collections across groups of libraries at different levels. In a major shift, a shared approach to print management is on the rise, and we anticipate that a large part of existing print collections, distributed across many libraries, will move into coordinated or shared management within a few years. This may involve physical

consolidation, or a more distributed approach where individual libraries declare commitments around parts of their collections. In this way, some attention shifts from the institution to supra-institutional structures as the venue for print collection management. Policy, organizational and service arrangements are now emerging around this trend.

The collective collection has been a major interest of OCLC Research. This is to be expected given the data we have in WorldCat about collective library holdings and OCLC's goal to make shared working among libraries more efficient. As interest in coordinated management of the collective print collection grows, we thought it was a useful time to pull together some of our writings on this topic in a single volume. This short piece provides some environmental introduction for the contributions that follow.

Interest in shared print strategies has had several drivers.

- **Google Books**. Google's December 2004 announcement of its intention to collaborate with five major research libraries to digitize their print collections and make them available for searching galvanized discussion about the collective print collection. Notably, it suddenly became possible to imagine the digitization of a large part of that collection, providing a significant alternative entry point to the print literature. The establishment of HathiTrust has provided community focus for curation of the digitized book corpus, even as the rate of digitization has slowed; it has moved curation to the network level. These developments have raised major questions about stewardship and permissible behaviors, questions that have resulted in legal actions. Institutions are beginning to plan their local collections in the context of the collective collection. For example, local decisions about print will increasingly be influenced by the emerging shared print management apparatus and by HathiTrust, as well as by the growing availability of e-books and more on-demand lending or acquisition practices.

- **The digital turn: changing patterns of research and learning**. While the print literature remains important to learning and research, overall use has declined to the extent that in some cases a misalignment is seen between current levels of investment in acquisition and management of print and the research and learning demands being placed on the library. The real cost of managing print has also become more apparent, at the same time as the use of digital materials increases. While print remains important in some contexts, there is a general move to digital resources, e-books and on-demand models.

- **Opportunity costs and space**. The opportunity cost of using space for the management of print collections is also becoming more apparent. Space is often required for higher value activities than storage of print collections, which are seen to be progressively releasing less value in actual use by students and faculty. Space is being reconfigured around broader education and research needs, and less around the

management of print collections. It supports social interaction around learning and research, and access to specialist expertise, equipment or communication facilities, as well as more exhibitions and interaction. In this context, many libraries are now actively managing down print.

- **Efficient access to print**. If fewer print materials are available in close proximity to users, it becomes important to ensure convenient discovery and delivery of those materials within new arrangements. Adequately supporting humanities scholars or other groups for whom print is important may depend on efficient delivery from within a system-wide apparatus of provision. For example, there are early discussions within the Committee on Institutional Cooperation (Big 10 institutions) about distributing print repositories to members specialized by subject, and ensuring rapid delivery across the system as a whole. For local service and political reasons it may be difficult to move in preferred directions without assurances about such broader provision.

- **A general move to collaboration**. Stronger models of collaboration are emerging, such as 2CUL[4] (Columbia and Cornell University Libraries) and the Orbis Cascade Alliance[5] (academic libraries in Oregon and Washington), which rebalance collections (as well as services, expertise and systems) in larger units.

Against this background it is interesting to consider this excerpt from a recent vision statement at the University of Arizona:

> Our goal is to be a primarily digital library. Simply put, it is no longer possible to sustain the massive print collections of the past. Our current physical plant is virtually full, and campus realities dictate that new buildings for the foreseeable future will be devoted to STEM initiatives with revenue generation potential. By shifting our focus from large print collections to electronic resources that are available anytime anywhere, the Libraries have moved from a "just in case" strategy to a "just in time" approach involving on-demand purchasing and large investments in speedy interlibrary loan. More than 94% of our serials and 20% of all our books are now electronic. In FY2012, our electronic book purchases exceeded 50% of total monographs purchased. In FY1999, we began to remove print materials duplicated by eresources to provide more out-of-classroom learning space for the campus. We have removed more than 750,000 print volumes to date. When print materials are desired or required, however, we are able to offer our quick and efficient interlibrary loan service or on-demand acquisition. We make collections decisions based on customer feedback, which has grown consistently more positive over the years.
>
> In 2005, we became the nation's first all-electronic Federal Government Depository Library. In 2011, we initiated on-demand, patron-driven access to electronic and print books. Listings for more than 60,000 scholarly books have been added to the library

> catalog; as these books get used by customers, we buy them and add them to our permanent collection. Research shows that patron-selected items get used more often, so this buying method maximizes the Libraries' purchasing dollars while giving users access to more information resources. (University of Arizona Libraries 2013, 4)

A system-wide perspective signals a real shift in emphasis. For most of its history, the library model was largely one of managing locally assembled collections. And the "goodness" of a library was strongly associated with its size, because more resources were available to its user base. This was natural in a print environment, where physical distribution in multiple collections in a just-in-case model was the most efficient way of meeting institutional needs. Responding on a case-by-case basis to satisfy student or faculty information needs as they were expressed would give rise to intolerable transaction costs.

This local assembly model was augmented by cooperation at the margins, whether through interlibrary loan or some cooperative approaches to collection development. As more of these print collections move into collective management, some core characteristics of the model change in interesting ways.

- **"Collection intelligence"** has to scale to the level of the system rather than the institution. In line with the local focus, libraries are used to treating their local catalogs as the definitive record of their collections, and shared cataloging and other approaches support this local management. As we think more about shared collections, this model flips. It becomes important to understand the characteristics of the collective collection so that local decisions can be made accordingly. The nature and quality of collective data becomes important to facilitate decision-making and effective processing. For example, while libraries now share data about the titles in their collections, they typically do not share data about individual copies. This becomes more important as libraries are interested in how many copies are in the system. OCLC's WorldCat and other union catalogs have become important tools in thinking about the characteristics of system-wide provision.

- **Preserving the scholarly record**. In the print world, preservation was a benign consequence of the redundancy inherent in the physical distribution model. Lots of copies, as they say, keep stuff safe. As this redundancy is reduced, a more planned or interventionist approach becomes important.

- **A balance of responsibilities.** Perceived responsibility for stewardship, provision or funding will vary across libraries in any evolving arrangement. Many research and national libraries recognize a mission-driven responsibility of stewardship to the scholarly and cultural record, and will undertake to work together and individually to discharge it as the environment changes. Other libraries may have specific regional or subject interests. However, many libraries may prefer to be consumers rather than

providers of shared collections, and may wish to participate more selectively, on a fee or membership basis, relying on collaborative or third-party arrangements to manage print collections. Others again may feel no need to make such a contribution. An important part of the shared print initiatives underway is to develop sustainability models that recognize the various interests at play within the system, and to put in place incentives to try to assure appropriate levels of participation.

- **Ownership.** It has been usual for libraries to think that they "own" the books in their collections. Google Books and HathiTrust have underlined that libraries actually have a bundle of rights—they can do some things but not others. At the same time, libraries are beginning to think about shared models of ownership and curation.

This is the context in which OCLC Research has developed a stronger interest in the contours of the collective collection. From our early work on the characteristics of the collections of the first libraries to participate in the Google Books program, there has been keen interest in knowing more about the composition of individual collections and about overlap and distinctiveness in the context of the aggregate library collection. We have looked at a variety of questions here. Our main resource has been WorldCat. At the time of writing, WorldCat contained approximately 300 million bibliographic records representing approximately 2 billion library holdings around the world.[6] While its coverage varies by type of library and region, WorldCat is the most complete record of global library holdings available.

We have three broad interests, which cluster around better understanding the existing collective collection and supporting the optimal evolution of reconfigured collections:

1. Understanding the characteristics of the collective print collection: how it is distributed across libraries and regions; its composition in terms of age, subject, copyright status and so on; levels of overlap, rarity and distinctiveness.

2. Supporting policy and service decision-making with good intelligence based on WorldCat and other data resources.

3. Understanding patterns or trends within the scholarly and cultural record. This is akin to "culturomics" (Michel et al. 2011) or "distant reading" agendas (Moretti 2013), which apply data-mining techniques to large aggregations of digitized text and metadata. It is a relatively new interest, and is not strongly represented in this volume. It is an area where we would like to encourage others to use WorldCat as a scholarly resource.

This report contains the following contributions:

Lavoie, Brian F. and Roger C. Schonfeld. 2006. "Books without Boundaries: A Brief Tour of the System-wide Print Book Collection." *Journal of Electronic Publishing* 9,2 (Summer). DOI: http://dx.doi.org/10.3998/3336451.0009.208.

> *Based on an analysis of WorldCat, this article discusses the characteristics of the North American print book "collective collection."*

Dempsey, Lorcan. 2006. "Libraries and the Long Tail: Some Thoughts about Libraries in a Network Age." *D-Lib Magazine*, 12,4 (April). http://www.dlib.org/dlib/april06/dempsey/04dempsey.html.

> *The long tail proposition is about how well supply and demand are matched in a network environment. This article considers library collections from this point of view and asks whether the current situation is the optimal system-wide arrangement of collections.*

Lavoie, Brian F., Lynn Silipigni Connaway, and Lorcan Dempsey. 2005. "Anatomy of aggregate collections the example of Google Print for libraries." *D-Lib Magazine*, 11,9 (September). http://www.dlib.org/dlib/september05/lavoie/09lavoie.html.

> *The initial Google digitization initiative galvanized interest in the composition and overlap of book collections. This important study looked at the overlap between collections of the original five library participants. It found that "rareness is common."*

Lavoie, Brian, and Lorcan Dempsey. 2009. "Beyond 1923: Characteristics of Potentially In-copyright Print Books in Library Collections." *D-Lib Magazine*, 15,11/12 (November/December). http://www.dlib.org/dlib/november09/lavoie/11lavoie.html.

> *Rights and allowable uses became a major area of discussion and contention around the emerging digitized corpus. This article aimed to provide some empirical basis for those discussions by exploring the characteristics of print books published in the US after 1923.*

Malpas, Constance. 2011. *Cloud-sourcing Research Collections: Managing Print in the Mass-digitized Library Environment.* Dublin, Ohio: OCLC Research. http://www.oclc.org/research/publications/library/2011/2011-01.pdf.

> *The objective of the project was to examine the feasibility of outsourcing management of low-use print books held in academic libraries to shared service providers, including large-scale print and digital repositories. It helped set the agenda around the emerging shared print discussions.*

Lavoie, Brian, and Günter Waibel. 2008. An Art Resource in New York: The Collective Collection of the NYARC Art Museum Libraries. Dublin, Ohio: OCLC Programs & Research. http://www.oclc.org/content/dam/research/publications/library/2008/2008-02.pdf.

> *This report examines the collective collection of four New York City-area art museum libraries, highlighting areas of distinctiveness and overlap that suggest opportunities for collaboration and collective action.*

Lavoie, Brian, Constance Malpas and JD Shipengrover. 2012. *Print Management at "Mega-scale": A Regional Perspective on Print Book Collections in North America.* Dublin, Ohio: OCLC Research. http://www.oclc.org/research/publications/library/2012/2012-05.pdf.

> *This report maps North American collections against mega-regions, areas which concentrate economic and social activity. It provides a new framework within which to think about organizational patterns of access and management, within a new geography of collections*.

Malpas, Constance. 2013. "Subsidence and Uplift—the Library Landscape." OCLC Research Hangingtogether.org Blog on 18 April. http://hangingtogether.org/?p=2680.

> *This short piece uses data about collection distinctiveness (in terms of subjects and names) to consider how HathiTrust is emerging as a significant center. This is one example of how ongoing reconfiguration will result in a rebalancing in how the collective collection is distributed across libraries, and shared print and digital repositories*.

Taken together, this work has helped shape the service and policy discussion as library collections are reconfigured by mass digitization and shared management initiatives. This theme is an important focus for us and we look forward to working with colleagues as the collective collection evolves in coming years.

Notes

1. See http://www.europeana.eu/portal/aboutus.html.
2. See http://dp.la/info/.
3. See http://www.cdlib.org/west/.
4. See http://2cul.org/node/17.
5. See http://www.orbiscascade.org/index/mission-statement.
6. See http://www.oclc.org/en-US/worldcat/catalog.html.

References

Michel, Jean-Babtist, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, John Orwant, Steven Pinker, Martin A. Nowak, Erez Lieberman Aiden. 2011. *Science*. 331 (6014): 176-182. http://www.sciencemag.org/content/331/6014/176.abstract#aff-10.

Moretti, Franco. 2013. *Distant Reading*. London: Verso. http://www.worldcat.org/title/distant-reading/oclc/813931586&referer=brief_results.

University of Arizona Libraries. 2013. *Everywhere You Are: The Library.* Arizona: University of Arizona Libraries. http://www.library.arizona.edu/sites/default/files/users/blakisto/vision.pdf.

Books without Boundaries:
A Brief Tour of the System-wide Print Book Collection

Brian F. Lavoie and Roger C. Schonfeld[1]

This paper was refereed by the Journal of Electronic Publishing's peer reviewers.

Abstract

Print book collections are facing significant transformation in response to mass digitization, remote storage, and preservation. These issues should be considered within a system-wide context in which individual print book collections are viewed not as isolated units, but rather as parts of a larger whole. As libraries look beyond the boundaries of their local print book collections to consider system-wide implications, they will need to be equipped with data and analysis about the system-wide print book collection. This paper provides a brief overview of the system-wide print book collection, defined as the combined print book holdings of libraries everywhere, as reflected in the WorldCat bibliographic database. Issues addressed include the size of the collection; holdings patterns; distribution by publication date and language; and the relationship of the system-wide print book collection to overall book production. The paper concludes with a brief discussion of some implications of the analysis, and possible directions for future research.

Print collections will likely undergo significant transformation as libraries continue to reshape themselves in the networked digital age. Some transformations will occur at the local level to meet the particular needs and requirements of a single institution and its users. However, it is likely that many more transformations will take place within a system-wide context, not individual library collections as isolated units, but rather as units of the aggregate library collection, the combined holdings of multiple libraries. Of course, the system can be defined at various levels of aggregation: by state, by region, nationwide, or even all libraries everywhere.[2] In all cases, however, the key point is that decisions regarding local collections will eventually and inevitably be taken with system-wide implications in mind.

This paper originally appeared in *Journal of Electronic Publishing* 9,2 (Summer).
DOI: http://dx.doi.org/10.3998/3336451.0009.208.

Today, decision making in a number of important areas would benefit from consideration of the system-wide context. Mass digitization programs such as the Google Print and Open Content Alliance projects raise fundamental questions about the size and scope of the entire system-wide collection. Given resource constraints, digitization efforts cannot hope to digitize every book in every library; some trade-offs will be necessary, such as digitizing selectively in certain disciplines, certain languages, or even certain libraries. Decision makers need to know the resources held in the system as a whole if they are to consider how multiple strategies might complement one another, avoid duplicative effort, and allocate resources in ways that maximize value and cost-effectiveness.[3]

A system-wide perspective is also useful in formulating collection management and preservation strategies. For example, retention, storage, and preservation decisions would benefit from knowledge of collection overlap across the system. Which resources are held redundantly by many libraries? Which resources are held by only a few libraries, or perhaps even a single library? Is the degree of overlap in the system acceptable to ensure resource survivability, given the risks of loss? Answers to these questions are necessary to inform the "weeding" of print collections, targeting scarce preservation funds to where they are needed most, and shared strategies for storage and preservation. Again, analysis of the system as a whole is necessary to support these decisions.

Digital and network technologies are breaking down the boundaries between local collections. Of course, this process has been at work for some time, as the increase in resource sharing has blurred the distinction between local and external collections. But now digitization, in combination with network connectivity, has accelerated the process, with one "copy" of a resource potentially being shared across many libraries. Digitization and online availability has opened up heretofore local collections to geographically dispersed audiences; therefore these technologies present new opportunities to expose users to resources beyond the local collection. Knowing what is in the system-wide collection, and how it is distributed across libraries, is an essential first step toward making that collection available to a system-wide audience.

In short, all of these factors (mass digitization, optimized collection management and preservation, and wide-spread access), as well as others, have contributed toward a shift in focus to the resources of the system, rather than individual library collections. But even as momentum in the library community has begun to shift towards a system perspective, the data needed to support and understand this perspective have not been widely available. In particular, data to support management and policy making at the system level, or at least with system implications in mind, are not routinely produced and analyzed. But this type of data will be increasingly critical as libraries and library collections become more deeply intertwined with the networked digital environment. System-wide analysis is certainly not a new concept; previous research studies have adopted, to a greater or lesser degree, a system perspective.[4] But in the networked age, the need to think "systematically" has never been greater.

In this paper, we focus on the collection of print books across libraries. Print books have been of particular interest recently, with the announcement of several mass digitization initiatives aimed at library print-book collections. This is not to say, of course, that print books are the only materials that would benefit from analysis at a system level. Given their rapidly progressing digital transformation, the serials literature (especially in the sciences) would provide another important terrain for system-wide analysis. However, analysis of the serials literature is sufficiently complex to warrant a separate study in its own right, and we therefore leave this for future work.

Questions addressed in this paper touch on some of the salient features of the system-wide print-book collection. How many titles does the system contain? What holdings patterns prevail within the system, especially in regard to the degree of overlap and the incidence of rare or unique materials? What are some of the characteristics of the print books in the system-wide collection, such as date of publication and language? This study offers only a brief sketch of the system-wide print book collection, with the objective of providing some examples of the kinds of data that could usefully be collected and applied. Ultimately, we hope this view of the system-wide print book collection helps libraries gain a fresh perspective on their collections, especially as they evaluate future needs and opportunities.

Defining the System

Any effort to count and characterize the components of a system naturally raises basic questions about how the system itself is defined. In regard to the system-wide print book collection that is the subject of this paper, the system in question should, ideally, consist of all print books held by libraries everywhere.[5] But assembling data on the system defined in this way presents practical difficulties that would be immensely difficult, if not impossible, to overcome. As a best approximation, we decided to define the boundaries of the system by the largest single source of cross-institutional bibliographic data available, OCLC's WorldCat database.

WorldCat is the world's largest and most comprehensive bibliographic resource. It currently contains more than 60 million bibliographic records and more than one billion holdings, reflecting the collections of more than 20,000 libraries worldwide. For the purposes of this paper, the print books represented in WorldCat serve as the system under study.

Defining a system in context of WorldCat has several limitations. All print books held by libraries have not been cataloged in WorldCat, nor have all libraries set their print book holdings in WorldCat. Moreover, WorldCat largely reflects North American library collections. But while WorldCat does not represent the entire universe of library collections, it provides the closest approximation to the ideal, and therefore is the best choice as a data source for a general overview of the system-wide collection of print books.

The statistics reported in this paper are based on a version of the WorldCat database from January 2005, containing about 55 million records and 950 million holdings.

The System-Wide Print Book Collection

The bibliographic records in WorldCat describe a wide variety of information resources, manifested in a range of formats. Figure 1 illustrates how the WorldCat database can be progressively filtered down to the subset of records describing print books.

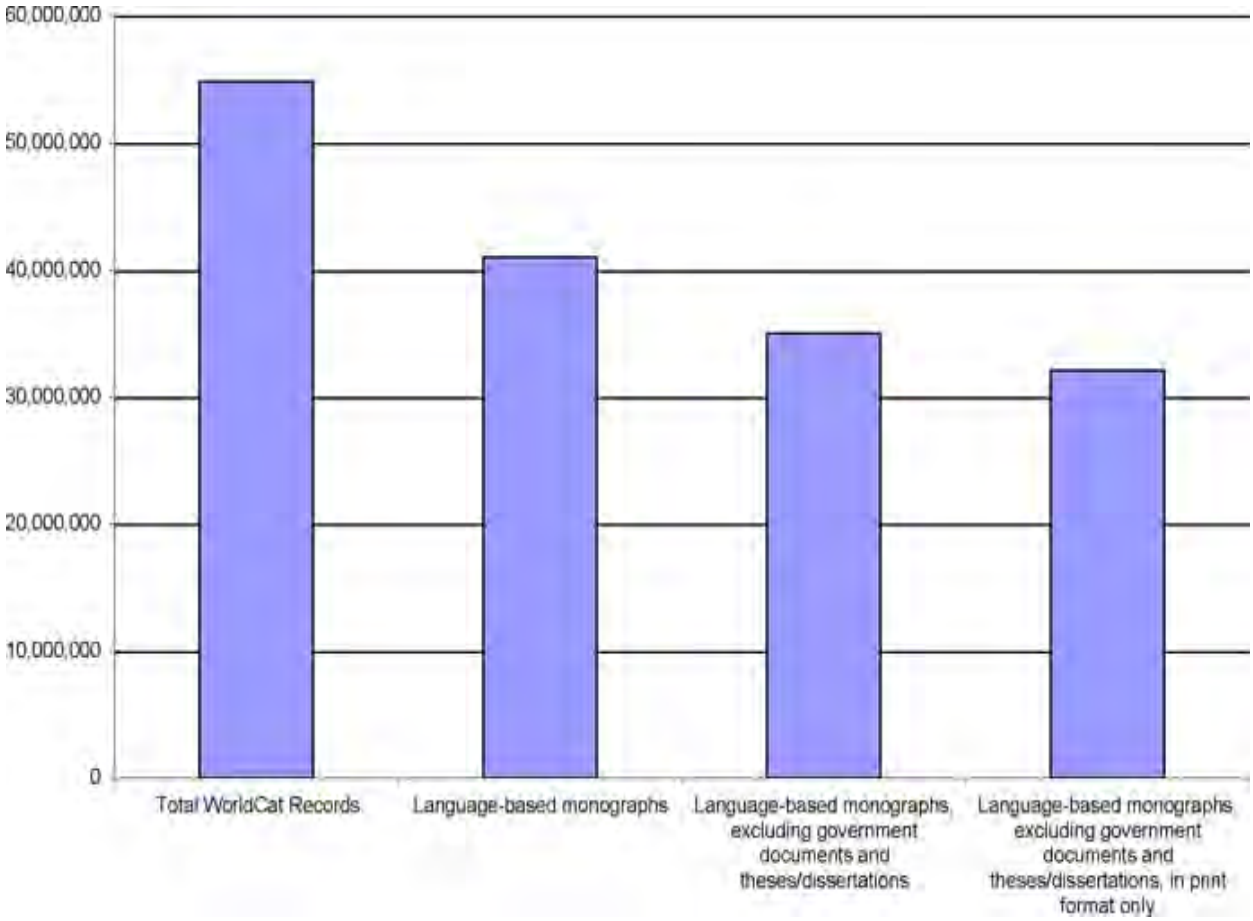


Figure 1. Print Books in World Cat

Of the approximately 55 million records in WorldCat as of January 2005, about 41 million described monographic language-based materials, which for the purposes of this study are considered "books."[6] Records describing theses or dissertations and government documents were then removed from this total; those materials are generally acquired and managed as separate segments of a library collection and therefore were excluded from the analysis. This lowered the total to about 35 million records. Finally, the scope of the analysis is limited to print books only, so all other formats, such as digital, microform, or Braille, were removed. This produced the final total of about 32 million print books cataloged in WorldCat.

The remainder of this paper discusses some of the salient characteristics of the system-wide collection represented by the 32 million print books cataloged in WorldCat.

Works and Manifestations

FRBR (Functional Requirements for Bibliographic Records) is a framework for understanding the relationships between various bibliographic entities, including works, expressions, manifestations, and items. Two bibliographic entities of interest for this study are works and manifestations. FRBR defines a work as "a distinct intellectual or artistic creation." Thus, Macbeth is a work. A manifestation, on the other hand, is a physical embodiment of an expression of a work. Thus, the Folger Shakespeare Library edition of Macbeth, published in paperback by Washington Square Press in 2004, is a manifestation of the work Macbeth. A single work can have multiple manifestations.

WorldCat records describe manifestations. Therefore, the 32 million print books cataloged in WorldCat represent 32 million distinct print book manifestations. Most of the analysis reported in this paper concerns manifestations, but it is also useful for some purposes to consider system-wide implications in terms of works. For example, works can shed additional light on questions of collection overlap beyond what is possible with manifestations alone. To explore some of these implications, the FRBR work set algorithm, developed by OCLC Research, was used to cluster the 32 million records in the system-wide print book collection into their associated works.[7]

There are a little over 26 million distinct works represented in the 32 million print book manifestations in the system-wide collection. This suggests that on average, each of these works contains just over one (actually, 1.2) print book manifestation, although certainly there is a subset of works containing many more.

By definition, each of the 26 million works associated with the system-wide print book collection contain at least one print book manifestation, but as figure 2 illustrates, less than half a percent contains both a print and a digital manifestation.

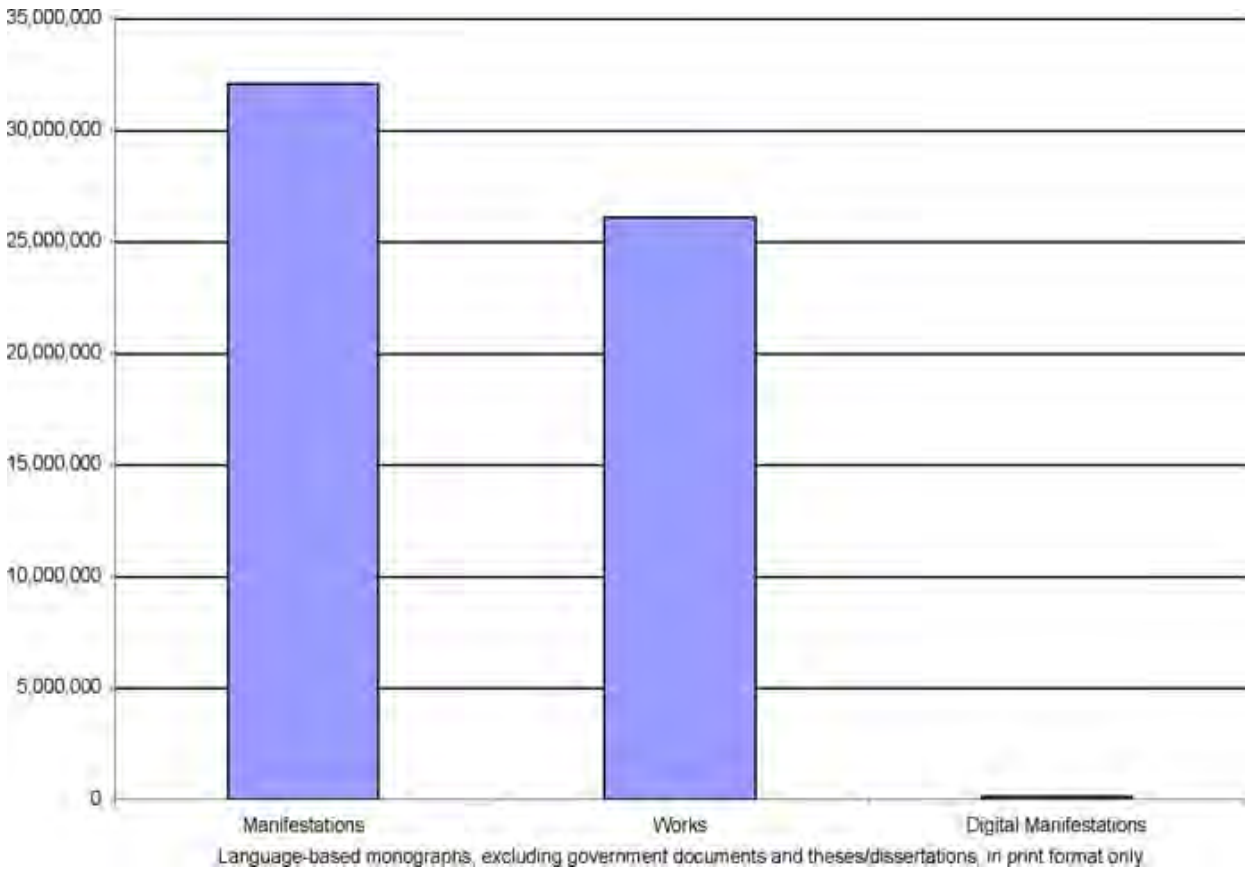


Figure 2. Print Manifestations, Print Works, and Digital Manifestations

This result must be interpreted carefully, because no one knows the exact proportion of digital resources held by libraries that are cataloged in WorldCat. But even if this figure is in reality 100 times greater, it is clear that only a small proportion of print books have been digitized. The transition of "legacy print content" to digital form has barely begun.

Holdings Patterns

The degree of collection overlap is a key issue in several areas, including digitization and preservation. Our examination of holdings patterns provides insight into the portions of the system-wide print book collection held redundantly by many libraries, and the portions held by only a few libraries, or perhaps even a single library.

Figure 3 illustrates some of the holdings patterns in the system-wide print book collection, illustrating the number of works held uniquely, those held twice, and those held more frequently throughout the system. By analyzing these statistics, we obtain a view of the maximum degree of overlap within the system-wide collection, since multiple manifestations of the same work are not regarded as distinct holdings.

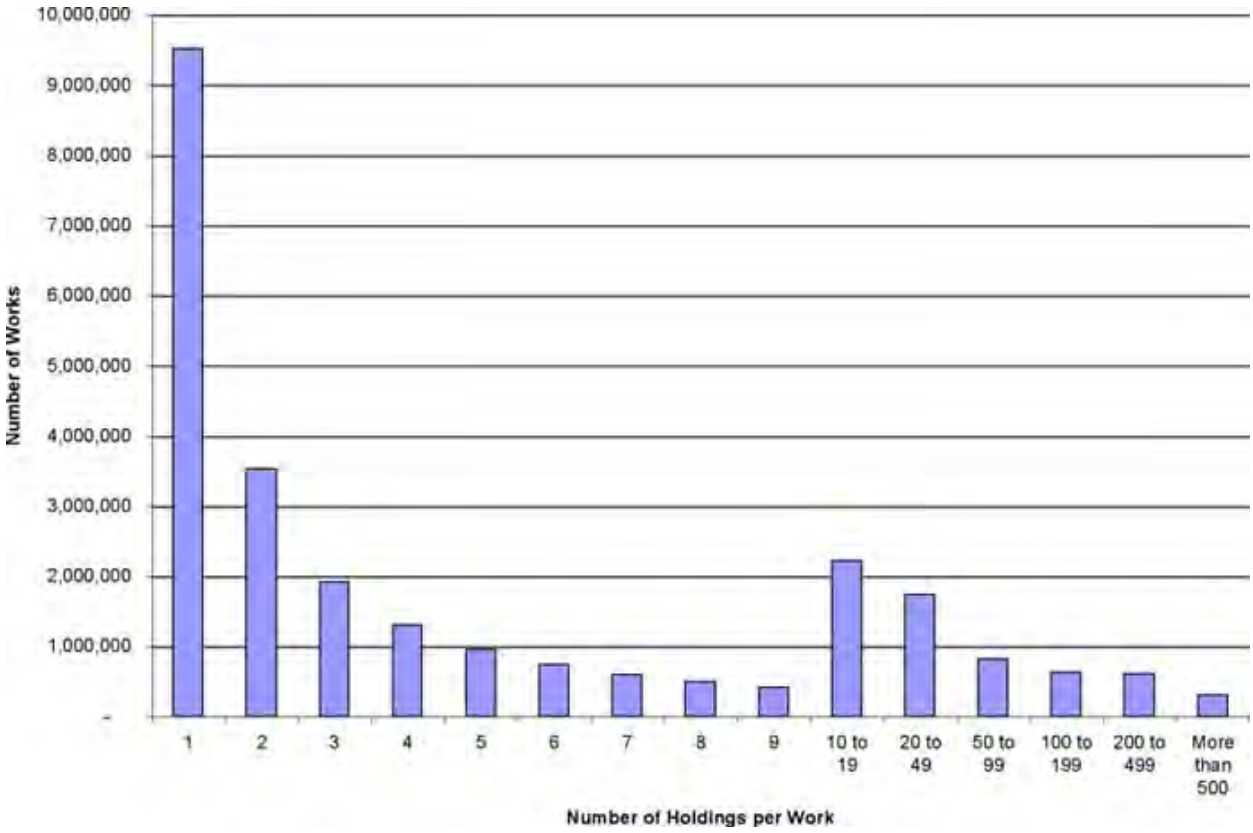


Figure 3. Books Held Multiple Times

Our data indicate the presence of about 9.5 million works that are held uniquely within the system, or about 36 percent of the total, and only approximately 2.4 million works with 50 or more holdings. While at first glance these figures may seem alarming, they deserve careful scrutiny. The framework for book survivability has generally relied on two components: the careful stewardship of rare books in segregated collections, and the overlap, or "preservation through proliferation," across the components of general circulating collections.[8] To be sure, if the 9.5 million works held uniquely are the circulating materials found in general collections, then something is amiss in the library system. If, on the other hand, these are largely rare books that are treated as such, then the situation appears far different.

We therefore examined a sample of 100 uniquely held works. We estimate that about 50% of these are in languages other than English which, as we will see momentarily, is not dramatically different from the overall share of books printed in languages other than English. Many of the English-language materials appear to be locally produced ephemera rather than traditional published books. Nevertheless, there are some items that would be recognized as traditionally published 20th century books that appear to be held uniquely within the system. The conclusions to be drawn from this sampling will vary for different institutions, but we are comfortable that they represent a fair view of the system's holdings.

Our sampling effort was designed to provide some context, but we believe that significant additional research is needed to understand uniquely held works. If, in fact, there are books that are not widely held but not treated as rare, a more systematic search for uniquely held, endangered books might be in order. The results of such a search could let us know how urgent it is to develop paper repositories to ensure that print works are not lost to the system. Such repositories also enable libraries to take more aggressive local approaches to collections management.[9]

On the other extreme, only 301,000 works are held 500 or more times—a relatively small share of the works in the system-wide collection. Because the system includes many public and school libraries along with academic and research libraries, it is all the more impressive that there are so few high-overlap works.

While our analysis is nothing more than a back-of-the-envelope assessment of collection overlap within the system, even a simple analysis such as this raises many questions that merit further research and policy debates.

Date of Publication

The rate of publication of print books has grown steadily over time. Figure 4 illustrates the distribution of books (manifestations) in the system-wide collection by year of publication. It is interesting to note the ebb and flow of book publication accompanying several important historical events, including a dramatic peak at the turn of the 20th century; troughs during the two World Wars and the Great Depression; and perhaps most importantly of all, the dramatic increase in publishing associated with the expansion of higher education and scientific research accompanying the start of the Cold War.

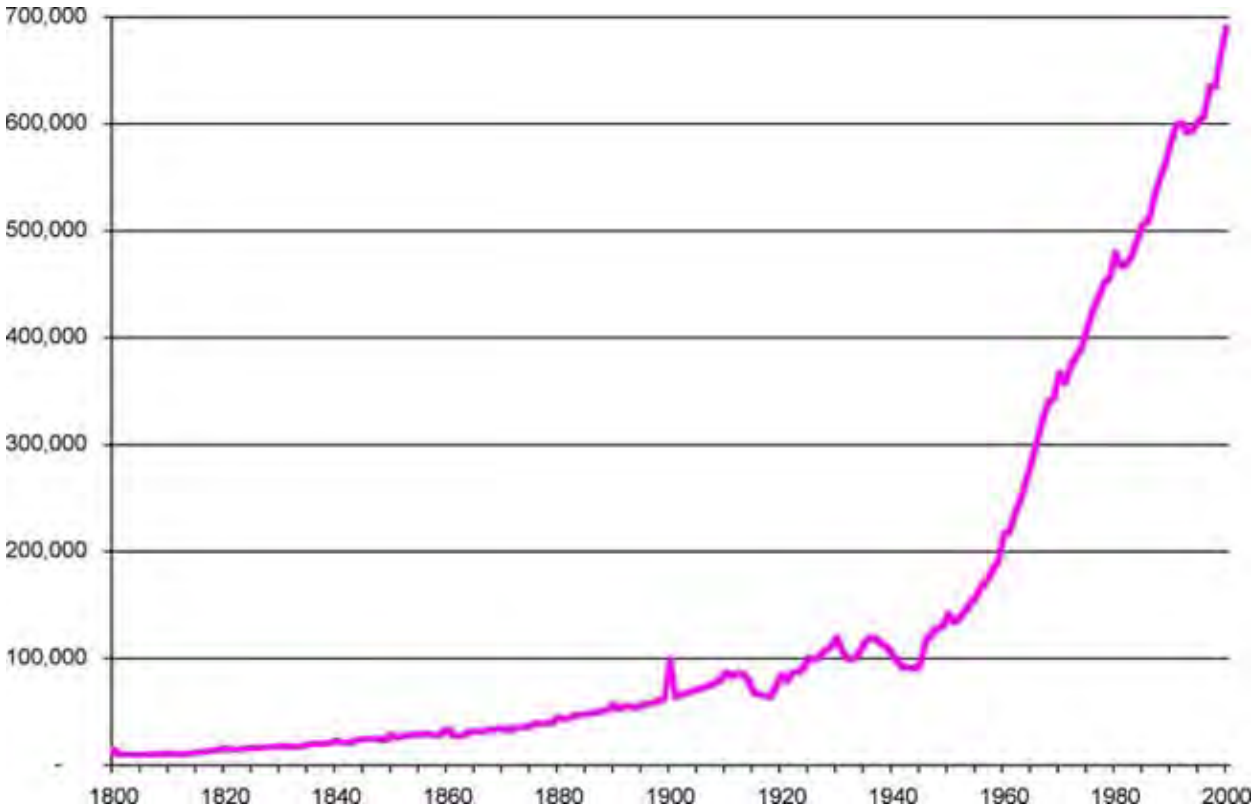


Figure 4. Print Manifestations by Year of Publication, 1800–2000

Cumulatively, the post-war increase in book publication is the dominating characteristic, as figure 4 illustrates. Approximately half of all books held in the system-wide collection were published after 1977. The share of these books published prior to 1923—a rough cut-off point for in-copyright vs. out-of-copyright materials, according to U.S. copyright law—is only 18%. Although the true share of out-of-copyright print books is undoubtedly higher than this due to nonrenewal of copyright for books published prior to the 1976 copyright law changes, the key point to be drawn from this figure is that a date-based approach to copyright permissions is not likely to yield a high proportion of books for mass digitization.

Language

There were approximately 450 languages represented in the system-wide print book collection. The distribution of these languages across the books in that collection is, of course, highly skewed. As figure 5 illustrates, a little more than half of the print books in the collection were published in English.

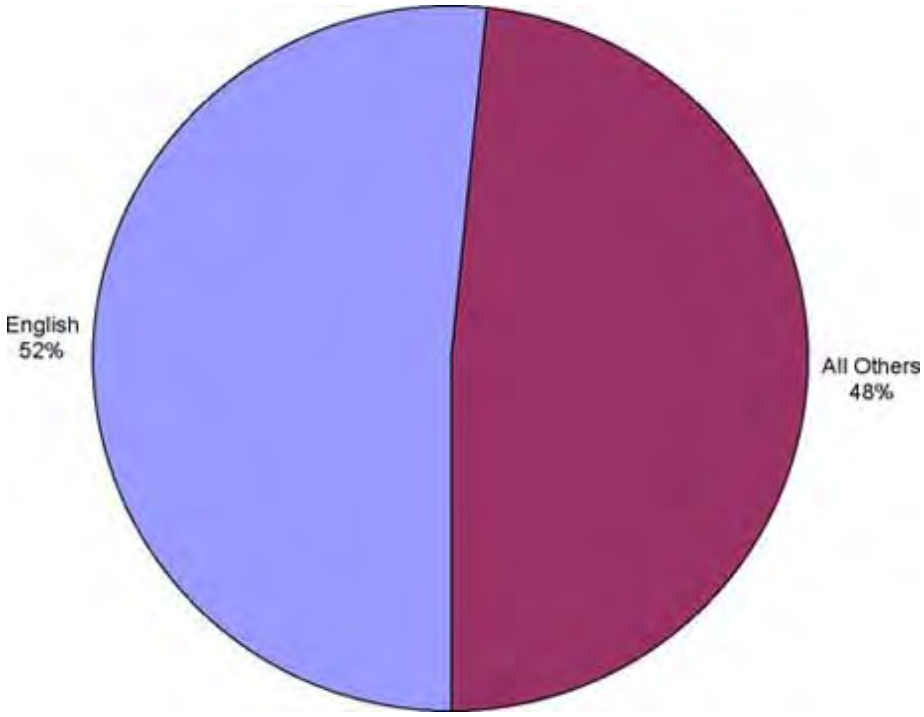


Figure 5. English Language

The incidence of all other languages in the system-wide collection is shown in figure 6. German, French, and Spanish are the most common languages after English; Chinese- and Japanese-language books make a surprisingly strong showing, probably reflecting at least in part the strength of area studies programs at some of the major research libraries. At the same time, we note the absence of any of the subcontinental languages among the top 25: Hindi, Urdu, Bengali, and Tamil all fall within the top 40, however, and in combination, would tie with Korean. This relative absence of subcontinental languages may be explained in part by the significant amount of English-language publishing in the region. And, in reference to the recent French concerns that US-based library digitization projects will omit French-language materials and thereby threaten the cultural balance of power,[10] we note that the system-wide collection—which, as noted above is heavily oriented toward North American libraries—contains more French-language books than any other language except English and German.

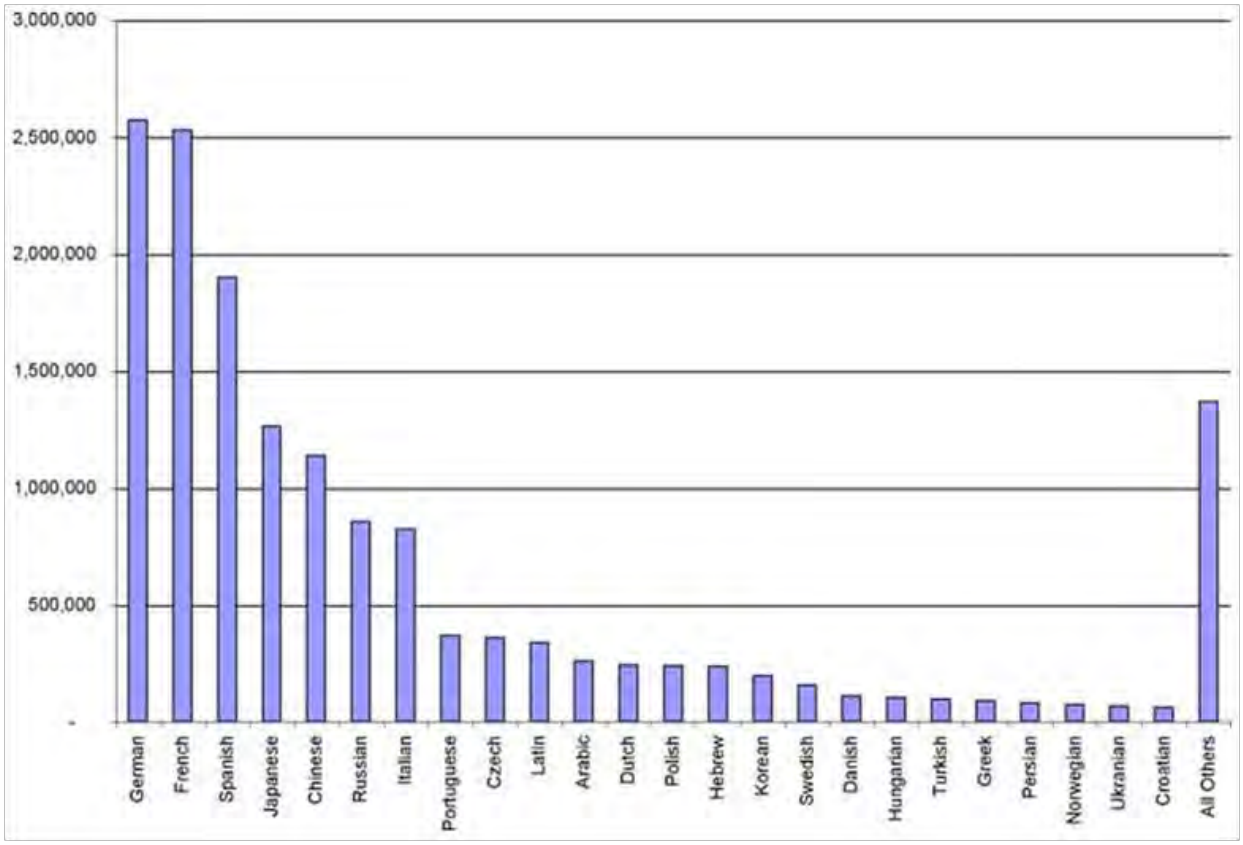


Figure 6. All Other Languages

Collections as a Share of Book Production

Embedded in these preliminary steps to characterize the system-wide collection are critical archiving questions. Fundamentally, we might ask, what does the system-wide collection not contain? What portion of our printed cultural heritage is unavailable? What portion has been lost? One way to approach these sensitive and complicated questions is to examine the share of total book production over time that is now a part of the system-wide collection.

This approach has several shortcomings. Most importantly, our data source incorporates many, but not all, of the collections held by libraries across the globe, which means it is likely that more book titles have survived and are available somewhere than our system-wide collection currently reflects. In addition, the data on total book production over time are estimates at best. Finally, the unavailability of a given book title need not imply that it should have remained available—the values that inform preservation choices, and the judgment calls needed to implement them, cannot be revealed through statistical methodology. Given these shortcomings, the estimates that we will present in this section raise questions that we believe merit further examination.

In the process of estimating both total book availability and book availability by year, we follow in the tradition of earlier researchers who were interested both in book production and its availability. Iwinski estimated that 10,378,365 books was the total historical book

production as of 1911.[11] His methodology in arriving at this estimate would probably have led him to undercount historical book production. By comparison, our figures show 4,568,987 print books with a publication date of 1911 or earlier. This could suggest that as many as 5.8 million book titles (all of which would be out of copyright today) may be absent from the system-wide collection.

In 1940, Merritt updated Iwinski's estimate, calculating that the historical total had grown to 15,277,276 published books, implying annual production since 1911 of about 156,000.[12] By comparison, our figures show 7,290,290 print books with a publication date of 1940 or earlier. That means the average number of print books with a publication date from 1912 through 1940 was 93,838. The book deficit by 1940 may have grown as high as 8 million; for the period 1912 to 1940 perhaps as many as 60,000 books per year did not, for one reason or another, enter the system-wide collection.

As figure 7 shows, although the number of books unavailable in the system increased with the publishing output over this 29-year period, the share of titles presently available actually grew somewhat.

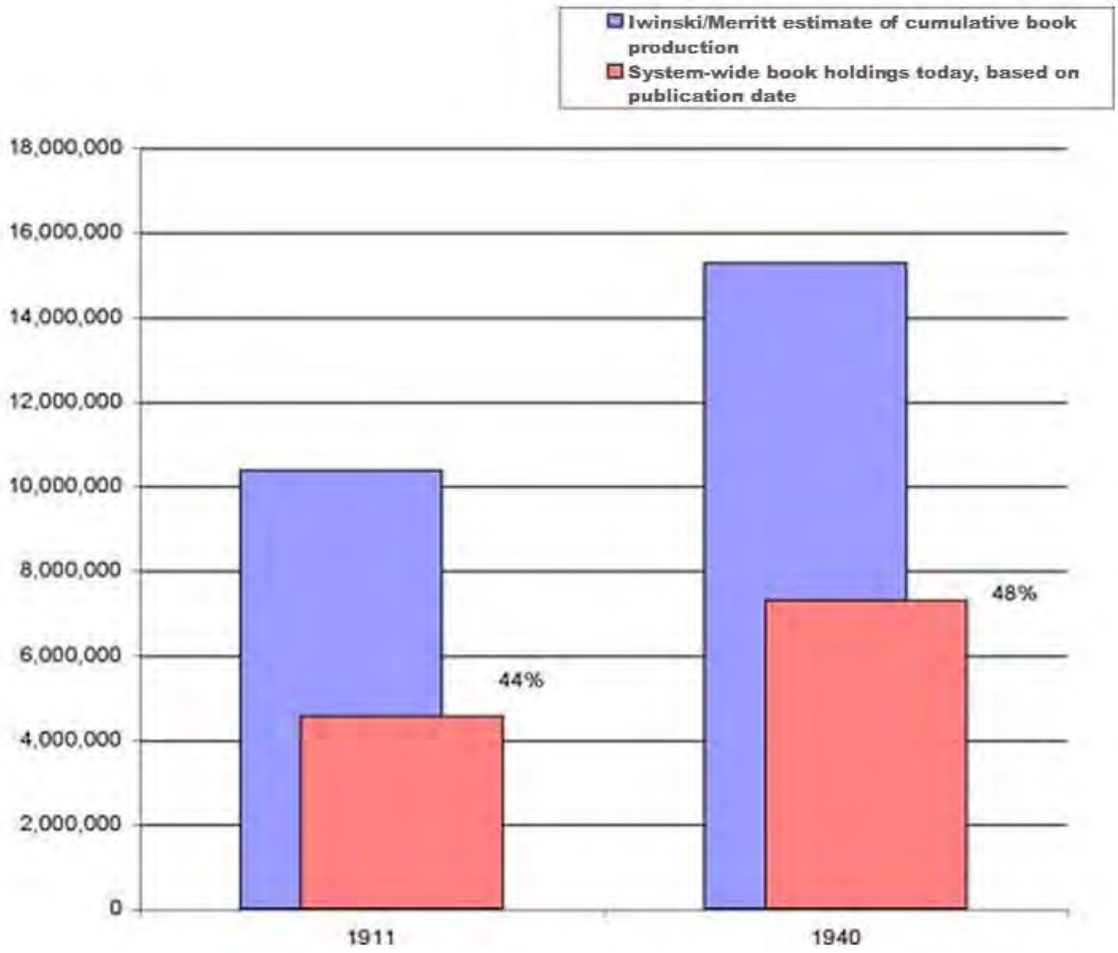


Figure 7. System-wide Book Gap, 1911 and 1940

The "book gap" illustrated in figure 7 is an extremely rough estimate; further work is needed to estimate this gap with more accuracy.

Although it would be highly desirable to perform similar calculations bringing us forward to the present, we have been unable to identify sufficiently reliable estimates of world book production for the latter half of the 20th century, and performing such an estimate was out of the scope of the present project. We can, however, use estimates of annual book production in recent years to see the share of present publications that are being collected. In recent years, UNESCO has attempted to calculate worldwide annual book production, with estimates in the range of 1 million book titles published per year.[13] In the system-wide collection, there are 689,496 books that were published in 2000; if book production in that year was of the magnitude estimated by UNESCO (roughly 1 million books) then the system is currently collecting about two-thirds of total book production.

Some Implications and Future Research Opportunities

Taking a system-wide view of library collections offers the opportunity to estimate the current size of our printed book heritage and to begin exploring some of its characteristics. It also suggests several important directions for further research and areas of caution for policy makers.

The public domain, relative to more recent publishing activity, is much smaller than many observers anticipated. There are important public policy implications to this finding, not least of which relate to digitization and "orphan works." It would be useful to understand better the characteristics of public-domain titles relative to those that remain under copyright: Are they more or less likely to be widely held? Does their country or language of publication differ substantially from in-copyright materials? What share of in-copyright books is out-of-print or "orphaned"?

We need to learn more about the characteristics of the rare and unique titles. Is the rareness that we identified a specter, brought about by cataloging shortcomings, or is rareness truly this pervasive within library collections? Can the rare materials be characterized in terms of subject matter, book type, and year, language, and location of publication? Are libraries holding these materials aware of their rareness? Are these books being adequately cared for?

Such analyses would shed light on some of the preservation issues raised in this paper, as well as provide a strong basis for policymaking. It would help us evaluate frameworks for dealing with print in an environment of large-scale digitization. As the first paper repositories are being developed, the library community should identify the optimal number of copies of a non-circulating book that should be preserved to guarantee its survival. This is fundamentally a risk-analysis question, and one that researchers can answer using tools such as actuarial analysis.

With roughly 32 million books in the system, mass digitization could create a collection that is significantly larger than our largest research libraries. Yet the fact that these titles are widely dispersed across the system presents significant organizational and information-sharing challenges to any mass digitization effort.

We also need to better analyze the holdings distribution across the system, especially in regard to rare or unique titles. Are the rare and unique titles concentrated in a set of major research libraries or distributed more broadly?

Some of our conclusions would be strengthened by comparative analyses on other union catalogs, especially overseas catalogs. This would allow a more detailed mapping of the system-wide print book collection, and raise new issues for analysis, especially about the "missing pieces." For new publications, what languages and regions are not collected commensurately with title output? Can we characterize the new publications that are not being collected? How do they differ from those that are typically accessioned? With more complete bibliographic data, is the "book gap" as large as it appeared to us?

Finally, it may be desirable to extend some of these techniques and recommendations to other formats beyond books. The system-wide serials collection, in particular, would merit study, although we believe the research task there to be much more complicated.

As the digital transformation reshapes the nature of print collections, these and many other issues will require the attention of librarians and other decision makers. As we learn how the system-wide collection contextualizes local collections, we might be able to develop new strategies for print collection management that reflect system-wide, rather than purely local, considerations. The observations and findings discussed in this paper are only a first step in this direction, but we hope they may set direction for discussions about the future of print books in the digital age.

Notes

1. Preliminary results from this study were presented at the 2005 CNI Spring Task Force Meeting. The presentation is available at: http://www.cni.org/tfms/2005a.spring/abstracts/presentations/CNI_schonfeld_system.ppt. The authors would like to thank Lorcan Dempsey, Joe Esposito, Kevin Guthrie, Joseph S. Meisel, Donald J. Waters, and conference/symposium participants at CNI, Ithaka, and the Yale Law School Library, as well as two anonymous reviewers, for helpful suggestions and comments.
2. There may also be cases when the system would be better understood as containing other institutional or individual repositories of texts and not just libraries as they are traditionally defined.
3. For example, OCLC researchers have examined the combined holdings of the five libraries participating in the Google Print for Libraries digitization initiative (see Lavoie, Dempsey and Connaway 2005).

4. See, in particular, Perrault's *Global Collective Resources: A Study of Monographic Bibliographic Records in WorldCat* (2002). Other important examples and reviews include:
   - Medina's "Duplication and Overlap among Library Collections: A Chronological Review of the Literature." (1995);
   - Perrault's "The Changing Print Resource Base of Academic Libraries in the United States." (1995);
   - Perrault's "National Collecting Trends: Collection Analysis Methods and Findings," (1999);
   - Perrault's "The Shrinking National Collection: A Study of the Effects of the Diversion of Funds from Monographs to Serials on the Monograph Collections of Research Libraries." (1994);
   - Potter's "Studies of Collection Overlap: A Literature Review." (1982).
5. For some purposes, of course, we might prefer to focus on the holdings of a type of library (such as academic libraries or national libraries), the libraries of a region or country, or other profiles.
6. Bibliographic criteria were byte 6 of the MARC21 leader equal to "a" (language material) and byte 7 of the leader equal to "m" (monograph). Language material restricts the focus to items that contain text, as opposed to musical scores, audiovisual material, objects, or maps. Monographs are items "either complete in one part … or intended to be completed in a finite number of separate parts," which excludes, among other things, serials. There is no authoritative bibliographic definition of a "book" available, but monographic language materials are often used to represent books. More information on these criteria are available at http://www.loc.gov/marc/bibliographic/ecbdldrd.html.
7. The OCLC Research work set algorithm was used to create a FRBRized version of WorldCat. The algorithm converts MARC21 bibliographic databases into FRBR work sets, where a work set is a cluster of all records (i.e., manifestations) pertaining to the same work. See http://www.oclc.org/research/projects/frbr/algorithm.htm.
8. The phrase was coined, and the strategy first identified as such, in Nichols and Smith's *The Evidence in Hand: Report of the Task Force on the Artifact in Library Collections* (2001).
9. For more background on the secure preservation framework that paper repositories are meant to afford, see Reilly Jr.'s *Developing Print Repositories: Models for Shared Preservation and Access* (2003).
10. See for example "The French Language: Google à la française," (2005)
11. See Iwinski 1911.
12. See Merritt 1942.
13. UNESCO Statistical Yearbook, 1999.

References

Iwinski, M.B. 1911. "La Statistique Internationale Des Imprimes." *Bulletin de l'Institut internationale de bibliographie* 16: 1-139.

Lavoie, Brian, Lorcan Dempsey, and Lynn Silipigni Connaway. 2005. "Anatomy of Aggregate Collections: The Example of Google Print for Libraries." *D-Lib Magazine* 11, no. 9. doi: 10.1045/september2005-lavoie.

Medina, Sue O. 1995. "Duplication and Overlap among Library Collections: A Chronological Review of the Literature." *Advances in Collection Development and Resource Management.* edited by Thomas W. Leonhardt, 1-60. Greenwich, Conn.: JAI Press.

Merritt, LeRoy C. 1942. "The Administrative, Fiscal, and Quantitative Aspects of the Regional Union Catalog." *Union Catalogs in the United States*, edited by Robert B. Downs, 1-128. Chicago: American Library Association.

Nichols, Stephen G., and Abby Smith. 2001. *The Evidence in Hand: Report of the Task Force on the Artifact in Library Collections*. Washington, DC: Council on Library and Information Resources.

Perrault, Anna H. 1995. "The Changing Print Resource Base of Academic Libraries in the United States. Collection Patterns from 1985 to 1989 in 72 ARL Libraries; Summary of a Ph.D. Dissertation; Presented at the 1995 ALISE Conference." *Journal of Education for Library and Information Science* 36: 295-308.

———. 2002. *Global Collective Resources: A Study of Monographic Bibliographic Records in WorldCat*. Dublin, Ohio: OCLC. http://www.oclc.org/research/grants/reports/perrault/intro.pdf.

———. 1999. "National Collecting Trends: Collection Analysis Methods and Findings." *Library & Information Science Research* 21, no. 1: 47-67. doi: 10.1016/S0740-8188(99)80005-X.

———. 1994. "The Shrinking National Collection: A Study of the Effects of the Diversion of Funds from Monographs to Serials on the Monograph Collections of Research Libraries." *Library Acquisitions* 18, no. 1: 3-22. doi: 10.1016/0364-6408(94)90067-1.

Potter, William G. 1982. "Studies of Collection Overlap: A Literature Review." *Library Research* 4: 2-21.

Reilly, Bernard F., Jr. 2003. *Developing Print Repositories: Models for Shared Preservation and Access*. Washington, DC: Council on Library and Information Resources.

"The French Language: Google à la française." 2005. *The Economist.* 31 March. http://www.economist.com/world/europe/displayStory.cfm?story_id=3819169.

Libraries and the Long Tail:
Some Thoughts about Libraries in a Network Age

Lorcan Dempsey

Discussions of the long tail that I have seen or heard in the library community strike me as somewhat partial. Much of that discussion is about how libraries contain deep and rich collections, and about how their system-wide aggregation represents a very long tail of scholarly and cultural materials (a system may be at the level of a consortium, or a state, or a country). However, I am not sure that we have absorbed the real relevance of the long tail argument, which is about how well supply and demand are matched in a network environment. It is not enough for materials to be present within the system: they have to be readily accessible ("every reader his or her book," in Ranganathan's terms), potentially interested readers have to be aware of them ("every book its reader"), and the system for matching supply and demand has to be efficient ("save the time of the user").[1]

Think of two numbers in this context. One is about interlibrary lending (the flow of materials between libraries), and the other is about circulation (the flow of materials within a library).

The first is that interlibrary loans (ILLs) account for 1.7% of overall library circulations. This goes up to 4.7% if we just look at academic libraries.[2] What this suggests is that we could do a better job making it easier to find and obtain materials of interest wherever they are, or, in other words, of aggregating system-wide supply. The flow of materials from one library to another is very low when compared to the overall flow of materials within libraries.

This might be what one would expect if the overlap between library collections were high. Last year, we at OCLC did some work looking at the aggregate collections of the Google 5 (G5) libraries. There we discovered that more than 60% of the G5 aggregate print book collection consists of books held by a single G5 library.[3] This suggests that collections are not as "vanilla" as is sometimes thought.

This paper originally appeared in *D-Lib Magazine*, 12,4 (April 2006).
http://www.dlib.org/dlib/april06/dempsey/04dempsey.html.

The second number is about circulation. We have also done some work looking at circulation data in two research libraries across several years. In each case, about 10% of books (we limited the investigation to English language books) accounted for about 90% of circulations. This shows that many books are not being borrowed (of course, some may be consulted in the library).[4]

These numbers suggest that many items in a specific collection may be underused, and that there is limited exchange of materials between collections. As we move forward, we will be increasingly asked if this is an optimal system-wide arrangement, especially as readers increasingly move to the network. We can think of requirements in the terms expressed by the following subset of Ranganathan's laws: books are for use; each book its reader; each reader his or her book; save the time of the user.

I want to look at some of these questions in more detail within a context established by the "long tail" discussion.[5]

The Long Tail

First a recap of the long tail argument, which since the publication of the original Chris Anderson *Wired Magazine* article has been much discussed.[6]

The argument is about how the Internet changes markets. In the "physical world," the costs of distribution, retail and consumption mean that an item has to generate enough sales to justify its use of scarce shelf, theatre or spectrum space. This leads to a limit on what is available through physical outlets and a corresponding limit on the selection potential of users. At the same time, the demand for a particular product or service is limited by the size of the population to which the physical location is accessible. This scarcity drives behaviors, about which we may have made mistaken assumptions:

> For too long we've been suffering the tyranny of lowest-common-denominator fare, subjected to brain-dead summer blockbusters and manufactured pop. Why? Economics. Many of our assumptions about popular taste are actually artifacts of poor supply-and-demand matching—a market response to inefficient distribution. [7]

These inefficiencies are mitigated in a network environment. And, accordingly, so the argument goes, we observe different behaviors with network services:

> Unlimited selection is revealing truths about what consumers want and how they want to get it in service after service, from DVDs at Netflix to music videos on Yahoo! Launch to songs in the iTunes Music Store and Rhapsody. People are going deep into the catalog, down the long, long list of available titles, far past what's available at Blockbuster Video, Tower Records, and Barnes & Noble. And the more they find, the more they like. As they wander further from the beaten path, they discover their taste is not as mainstream as they thought (or as they had been led to believe by marketing, a lack of alternatives, and a hit-driven culture).[8]

Netflix, for example, aggregates supply as discussed here. It makes the long tail available for inspection. However, importantly, it also aggregates demand: a larger pool of potential users is available to inspect any particular item, increasing the chances that it will be borrowed by somebody.

Anderson provided some interesting numbers to show the impact of this phenomenon in his original article, and these have been updated on his website.[9] He notes that the aggregation of the long tail is a major part of the business model of the leading Internet services (Amazon, eBay, Google, etc.). Google, for example, services the long tail of advertising—those for whom the bar was too high in earlier times of scarce column inches or broadcast minutes. And by aggregating demand, delivering a large volume of users, they increase the chances of the advertisement being seen by somebody to whom it is relevant.

Of course, merely being on the web is only a part of the issue. What the web allows is consolidation. Anderson's examples are massive, consolidated web presences. As suggested a moment ago, this consolidation has two aspects: aggregation of supply and aggregation of demand. Each is important.

Five things come to mind about the aggregation of supply and demand. The first is transaction costs, the costs incurred—whether in attention, money, expertise or some other resource—in achieving one's goal. High transaction costs inhibit use: they increase the friction in the system; low transaction costs encourage use: they increase flow through the system. iTunes for example, has low transaction costs. The burden of discovering tracks of interest, transacting for their use and downloading them is low. The tracks are immediately available. Netflix has higher transaction costs given the delays caused in the mail system, but still works to provide as frictionless a workflow as possible for the user. We can think of two aspects of transaction costs: search costs and fulfillment costs. How difficult is it to discover something, and once it is found, how difficult is it to acquire a service or an object?

The second thing to come to mind is the availability of consolidated data about choices and behaviors.[10] Netflix, Amazon, Rhapsody and others refine their service based on what they know of their users' choices, mined directly from the aggregated clickstream. This allows them to develop services that can further develop reflexively based on usage, and that can be tailored around particular behaviors and preferences. Furthermore, additional services can by built by leveraging this mined user data—recommender services for example. These services potentially reduce transaction costs, because they use aggregate data about behaviors to better target their offerings.

The third thing to consider is inventory. These large web presences consolidate inventory: they are not encumbered by the costs of massively redundant, just-in-case inventory, scattered through multiple physical delivery points. This consolidation may happen in virtue of the digital nature of the collections, as with iTunes. Or, where physical inventory is

involved, as with Amazon, they can consolidate in strategic locations, or with particular suppliers, as inventory need not be tied to physical storefronts. They manifest their store

through the management and presentation of data, not through the actual display of goods in a physical store. And, of course, consolidation of inventory may reduce transaction costs by streamlining fulfillment.

The fourth thing is about navigating the consolidated resource. Google introduced a major innovation with its ranking approach, by aggregating and mining the linking choices made by web page authors. Amazon is interested in rich interconnection through reviews, wish lists, reader selected lists, the various "phrases" (capitalized and statistically improbable), and so on. Amazon provides a rich texture of suggestion.[11] In each case, simple aggregation is not good enough: also needed are effective ranking, recommending, and relating.

And finally, large web presences help aggregate demand. The level of use of a resource partly depends on the size of the population to which it is accessible. One aspect of the long tail argument is that the aggregation of demand—extending the population to which a resource is accessible—means that resources have a better chance of finding interested users. In other words, use will extend down the long tail. So, as discussed above, Netflix finds viewers for movies that might not move in a physical outlet, because Netflix aggregates demand across a larger population than a single physical store can. Google, iTunes, Amazon, eBay: the gravitational pull of these resources on the open web means that they have achieved a wide audience of potential buyers or sellers. This increases the chances that resources they disclose will rendezvous with interested consumers on the web. So, they aggregate demand by drawing users to them. However, increasingly they also go to users. Google, Amazon and eBay, for example, are very actively trying to reach into multiple user environments through the use of toolbars, APIs and other approaches.

Libraries and the Long Tail

So, now let's turn back to libraries, and focus on these two issues: the aggregation of supply, and the aggregation of demand. For convenience of discussion, I focus primarily on books, drawing in other resources occasionally. I hope readers can see how the discussion can be extended to cover other parts of collections.

The Aggregation of Supply and Demand in Libraries

Libraries have been subject to the same physical constraints as, say, bookstores, albeit within a different service context. The library collection is not limited to the current or the popular: the library has some responsibility to the historical record, to the full range of what *has been* made available as well as to what is *now* available. That responsibility varies by library type, and is variably exercised. The library has met that responsibility in two ways: by assembling a

local collection, and by participating in systems of extra- and inter-library provision. These latter systems may be organized in different ways; the resource-sharing consortium is a common pattern, and a library may belong to several.

The library collection is driven by local perception of need and available resources: collection development activities exist to balance resource and need. A large research library and a busy public library system will have different profiles, but both are influenced by physical constraint. In the material world, the transaction costs of access to a distributed library collection are high, so those libraries that could afford it sought to amass large local collections in order to aggregate supply locally. Think of, for example, the large just-in-case research library collections. And, indeed, we are still measuring research library strengths by number of volumes. A busy public library may move towards the bookstore model. I was at a presentation recently about a busy public library system in an affluent suburban area. They turned over 15% of their stock per annum: they want stock to circulate and to keep it fresh for a demanding audience; just as in a bookstore, titles had to justify their occupation of limited shelf space.

Next I discuss the issues I mentioned above (transaction costs, data about choices and behaviors, inventory, navigation and aggregation of demand through major web presences) as they apply to libraries.

Transaction Costs

A library user has a range of discovery tools and services that provide access to a fuller range of scholarly and learning materials. This in turn is supported by a well-developed apparatus of deposit libraries, resource sharing systems, union catalogs, cooperative collection development, document supply, and other collaborative and commercial services. This "apparatus" may be imperfectly and intermittently articulated, but it is a significant achievement nonetheless. What an individual library may not be able to supply should be available within the overall system in which libraries participate.

However, this availability is bought at the expense of some complexity, which in turn means that the transaction costs of using the system are high enough that some needs go unrecognized or unmet. A library user may not be familiar with available tools or may not be aware that other materials are available. Local policies may restrict some types of access. Thus, historically, one can say that while library services explicitly aim to aggregate supply and demand both to meet user needs and to maximize use of resources within an overall apparatus of provision (see Ranganathan's laws again), imperfect articulation of that apparatus means that users are variably served. To make this more concrete think about the D2D chain: discover, locate, request, deliver[12]. Here lack of integration increases transaction costs. By integration, I mean within processes (there are many discovery options, for example) and between processes (the processes are not always connected in well-seamed ways).

- **Discover.** The discovery experience is a fragmented one. A user has a range of discovery tools available and may not always know which is the most suitable. This is especially the case with the journal literature, in which case the deployment of metasearch approaches is a partial response. Even for books, users may have to navigate a patchwork of catalogs to find what they are looking for; search costs are high.[13] What might one do? One approach is consolidation: fewer but larger pools of metadata to support discovery would help. Another is "syndication," moving the metadata to where it might more readily rendezvous with the reader. I use syndication as a general term to include such ideas as letting metadata flow into citation managers, search engines and other resources, and to expose it in services upon which other applications may build. The latter is familiar to us from Amazon, which can make its data and services available in other interfaces through its APIs.

- **Locate**. Having identified an item of interest, a user needs to find a service that will supply it. This may be as simple as noting a call number and walking to a shelf. Or it may involve a resolution service that actually provides several service options. Or it may involve a further discovery experience in a library resource if the item was originally found outside the library. This latter case is especially interesting, as library users have many more discovery options outside the library than within it. What is needed is a way of connecting the discovery experience to a library service. Here Coins provides a potential approach, coupled with various browser tools.[14]

- **Request.** This is another transaction, which may involve one or more steps. It can be simple, as in placing a hold, or more complex if a form has to be filled out, and so on. Increasingly, libraries may want to route requests in several directions: allowing a user to buy from Amazon, initiate an ILL request, initiate a document supply request, or place a hold on the requested material.

- **Deliver.** Again, several potential options exist for resource delivery, which can involve more or less difficulty depending on how the delivery options are presented and on the disposition of supplier and user. This ties interestingly to the inventory question, and I come back to this below.

You get the idea: at each stage, there are potentially many processes that need to be connected, and they potentially need to be connected to each other in different combinations. The better connected, the lower the transaction costs. Indeed, it is interesting to wonder if resolution services will move more to the center of library operations, as they are effectively "service routers" connecting multiple discovery experiences to multiple fulfillment services.

Data about Choice and Behaviors

Transactional and behavioral data is used to adapt and improve systems. In the library community we have not yet fully exploited these opportunities. Examples of such data are holdings data (choices made by libraries), circulation and ILL data (choices made by users), and database usage data (choices made by users). We have few services yet that aggregate such data. Libraries are increasingly interested in using this data to refine services and build new services as discussed above. Think of recommender services based on circulation data, for example. As new services and user behaviors co-evolve in changing digital spaces, it is likely that we will want to capture new forms of data.

Inventory

The historic library model has been physical distribution of materials to multiple locations so that the materials can be close to the point of need (as in the bookstore model). And again, in the network environment, of course, this model changes. Resources do not need to be distributed in advance of need; they can be held in consolidated stores, which, even with replication, do not require the physical buildings we now have. As we move forward, and as more materials are available electronically, we will see more interest in managing the print collection in a less costly way. We can see some of this discussion starting in relation to the mass digitization projects and the heightened interest in off-site storage solutions. In each case, there is a growing interest in being able to make investment choices that maximize impact—based, for example, on a better understanding of what is rare or common within the system as a whole, on what levels of use are made of materials, and so on. In fact, again looking forward some time, it would be good to have management support systems in place that make recommendations for moving to storage or digitization based on patterns of use, distribution across libraries, and an agreed policy framework.

There are two medium-term questions that are of great interest here. First, what future patterns of storage and delivery are optimal within a system (again, where a system may be a large library system, a state, a consortium, a country)? Think of arranging a system of repositories so that they are adjacent to good transport links for example, collectively contracting with a delivery provider, and having better data intelligence for populating the repositories, based on patterns of use and demand. Second, think of preservation. Currently, we worry about the unknown long-term costs of digital preservation. However, what about the long-term costs of print preservation? I contend that for many libraries they will become unsustainable. If the use of large just-in-case collections declines, if the use of digital resources continues to rise, if mass digitization projects continue, then it becomes increasingly hard to justify the massive expense of maintaining multiple collections—especially where there is growing demand for scarce space. Long-term we may see a shift of cost from print to digital, but this can only be done if the costs of managing print can be reduced, which in turn means some consolidation of print collections.

As these questions push us towards a system-wide perspective, aggregating data about supply and demand gives a better sense of what is collectively held in the system, what is collectively being used in the system, and from this, how decisions about the optimum disposition of collections can be facilitated.

Navigation

Library aggregations have not exploited the structure of the data very effectively to support navigation. The interest in faceted browse, FRBR,[15] recommendation, ranking by holdings or other data, and so on is testament to a realization that better ways to exploit the large bibliographic resource are needed. Ranking, recommending, and relating help connect readers to relevant materials and also help connect the more heavily used materials to potentially useful, but less used, materials.

Aggregation of Demand

The library resource is fragmented. It is fragmented within the library (there are many databases to choose from; they may be organized in a different ways in different libraries). It is fragmented across libraries, as discussed above. In the new network environment, this fragmentation reduces gravitational pull. It means that resources are prospected by the persistent or knowledgeable user, but they may not be reached by others to whom, nevertheless, the resources are potentially useful. Additionally, the library resource cannot be very well assimilated into user workflows. The availability of RSS feeds, APIs, and other approaches are making it possible to insert the library into the user environment (rather than always expecting the user to come to the library environment), but we are only in early stages in this regard. There are two issues here. The first is that libraries may need to do more work to aggregate demand within their own institutions. And one approach to this is to consolidate the library web presence (think of metasearch for example) and to project library services into user workspaces (embedding database searches in course pages, for example). The second issue is that it may be difficult for individual libraries to aggregate demand above the individual library level. Union catalogs and resource sharing systems have historically operated above an individual library level, and we are now seeing organizations who supply those services thinking about how to redevelop as major web presences that help aggregate demand (backed up by aggregated supply). Examples here are RedLightGreen, OpenWorldCat, and Libraries Australia. Library organizations are also very keen to be visible within the major web-based search engines and bookselling sites. Of course, one way for a library to try to reach its local audience is to make its resources visible in these major web presences, which is where its users spend much of their time and attention.

This provides an interesting perspective from which to view Google Scholar and Google Book Search, in particular their interaction with libraries. Take Google Book Search: what Google is

doing here is potentially aggregating demand for books: it will be interesting to see what influence this has on their use. Presumably a case has been made that there is potential interest in the full scope of those collections, or, in other words, in moving down the long bibliographic tail (and remember the figures I presented above about the current situation). They are also aggregating demand for books and journals through Google Scholar. And, to avoid frustrating users, they are aggregating supply behind the discovery experience. Hence they are working with resolver data and multiple suppliers to complete the locate/request/deliver chain for journal materials. In addition, they are working with OCLC to connect the Google Scholar discovery experience to the "Find in a Library" option for fulfillment. What OCLC is doing is making metadata about those books available to the major search engines and routing users back to library services, to complete the D2D chain for books. To the extent that a large amount of materials are made available through these services, Google is aggregating demand, aggregating supply, and reducing transaction costs.

Logistics and Libraries

So, briefly, what are some consequences for libraries?

Libraries have rich, deep collections, and the aggregate library system is a major achievement. However, in our current network environment, libraries compete for scarce attention. This suggests that if the "library long tail" is to be effectively prospected then the "cost" of discovering and using library collections and services needs to be as low as possible.

This is a logistics issue. Logistics is about matching supply and demand in a timely fashion across a network of potentially many parties. Within a particular domain, this is what libraries have always done, and some of the recent innovation in libraries has been precisely to automate supply chains (think of resolution services, for example).

Here are some ways of improving aggregation of supply and demand:

- Unify discovery experiences: Fragmentation is costly, and fewer but larger resources might help.
- Project library discovery experience into other environments: search engines, browser tools, RSS aggregators, etc.
- Better integrate D2D, both within operation (for example, combine request options—Amazon, place hold, ILL, ...) and between operations: The aim should be to be able to place a "get it" button anywhere and guide the user through simple choices.
- In the medium term, explore how "inventory" and "distribution" are managed across a system: (This should be done whether a system is a library, a consortium, a state, or a country).

- Utilize better "intelligence" within the network: This involves better representing the entities within the network. It touches on the growing interest in "registries"—registries of services (a registry of deep OPAC links, or OpenURL resolvers, or Z39.50 targets are examples here), registries of collections (a registry of database descriptions is an example), registries of institutions (see the very fine National Library of Australia Libraries Gateway for example), registries of policies (increasingly important, as libraries will organize within policy frameworks), and so on. In this context, it is interesting to reflect that the distinctive value of union catalogs is the holdings data: a union catalog is a registry of "information object" data related to holding institutions. Collectively, the registry data discussed here will drive the applications that support "library logistics."

- Provide transaction support: In an environment of multiple transactions between libraries it is useful to have a way of tracking and reconciling between libraries. OCLC's Fee Management service[16] is an example of a service that supports some classes of transaction. (Think of how PayPal has released various possibilities of interaction.)

- Aggregate demand through significant web presences: If more users are exposed to library collections, the collections will be used more. Of course, in some contexts demand from external users has been one reason for not more widely exposing collection information. However, the dynamics of the network have changed use. The major Internet search presences are often the first and last resorts of research, and fragmentation of library resources reduces their gravitational pull. Libraries are having to compete for the attention of their own users. They need to be in user environments, and the open web is now very much part of those environments. This leads to consideration of the discovery strategies mentioned.

Conclusion

Libraries collectively manage a long tail of research, learning and cultural materials. However, we need to do more work to make sure that this long tail is directly available to improve the work and lives of our users. Books, after all, are for use.

I mentioned Ranganathan near the beginning of this article. Ranganathan's five "laws" have classic status in the library community. They express something that remains relevant even as contexts change. Think of "book" as shorthand for the range of resources the library provides.

I wrote about the "long tail" in terms of aggregation of supply and aggregation of demand. In this context, aggregation of supply is about improving discovery and reducing transaction costs. It is about making it much easier to allow a reader to find it and get it, whatever "it" is. Or, in other words, "every reader his or her book." Aggregation of demand is about mobilizing

a community of users so that the chances of rendezvous between a resource and an interested user are increased. Or, in other words, "every book its reader." Finding better ways to match supply and demand in the open network will "save the time of the user."

How we do this is a part of a general reshaping of activities and organizations in a network environment. We need new services that operate at the network level, above the level of individual libraries. These may consolidate D2D, or management of collections, or other services. They may be collaboratively sourced or provided by third parties. It does pose interesting questions about how resources are allocated to best achieve local impact and system-wide efficiencies. This change also shows that the library continues to be "a growing organism."

Acknowledgment

Discussion with my colleague Brian Lavoie improved this article; I remain responsible for its deficiencies.

Notes and References

1. Ranganathan's Five Laws of Library Science remain a valuable touchstone http://en.wikipedia.org/wiki/Five_laws_of_library_science. They are listed there as:
    1. Books are for use.
    2. Every reader has his or her book.
    3. Every book has its reader.
    4. Save the time of the reader.
    5. The library is a growing organism.
2. The source of this number is OCLC marketing, based on available data.
3. Brian Lavoie, Lynn Silipigni Connaway and Lorcan Dempsey. Anatomy of Aggregate Collections: The Example of Google Print for Libraries. *D-Lib Magazine*, Vol. 11, No. 9, September 2005. http://dx.doi.org/10.1045/september2005-lavoie. (Reprinted in *Zeitschrift fur Bibliothekswesen und Bibliographie*, Vol. 52, No. 6, 2005. pp 299-310).
4. This data is from unpublished work by Lynn Silipigni Connaway and Edward T. O'Neill.
5. This short article adapts an earlier piece: Lorcan Dempsey's Weblog: Dempsey, Lorcan. 2006. "Libraries, Logistics and the Long Tail." February 15. http://orweblog.oclc.org/archives/000949.html. Some responses to that post are discussed in Lorcan Dempsey's Weblog: Dempsey, Lorcan. 2006. "Systemwide Activities and the Long Tail." http://orweblog.oclc.org/archives/000955.html.
6. Anderson, Chris. "The Long Tail." *Wired Magazine.* Issue 12.10 October 2004. http://www.wired.com/wired/archive/12.10/tail.html.
7. See note 6 above.
8. See note 6 above.
9. See Chris Anderson's long tail website at http://www.thelongtail.com.

10. Usage data seems too flat an expression for what I mean here. Elsewhere I have used the phrase "intentional data," modeled after John Battelle's characterization of the "database of intentions." This is the accumulated usage data of the Internet search engines. See: Battelle, John. 2005. *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*. New York: Portfolio, p. 6.
11. Lorcan Dempsey's Weblog. 2006. "The Simple Search Box And The Rich Texture Of Suggestion." 12 March. http://orweblog.oclc.org/archives/000966.html.
12. Lorcan Dempsey's Weblog. 2005. "Discovery, Locate ... Horizontal And Vertical Integration." November 20, http://orweblog.oclc.org/archives/000865.html.
13. See the article by Swarthmore faculty member, Burke, Tim. 2004. "Burn the Catalog." 20 January. http://www.swarthmore.edu/SocSci/tburke1/perma12004.html.
14. Hellman, Eric. "OpenURL COinS: A Convention to Embed Bibliographic Metadata in HTML." Last updated 16 June 2009. http://ocoins.info/.
15. FRBR (Functional Requirements for Bibliographic Records). See: What is FRBR? http://www.loc.gov/cds/FRBR.html.
16. OCLC's Fee Management service, http://www.oclc.org/resourcesharing/features/feemanagement/.htm.

Anatomy of Aggregate Collections: The Example of Google Print for Libraries

Brian Lavoie, Lynn Silipigni Connaway, Lorcan Dempsey

Introduction

Google's December 2004 announcement[1] of its intention to collaborate with five major research libraries—Harvard University, the University of Michigan, Stanford University, the University of Oxford, and the New York Public Library—to digitize and surface their print book collections in the Google searching universe has, predictably, stirred conflicting opinion, with some viewing the project as a welcome opportunity to enhance the visibility of library collections in new environments, and others wary of Google's prospective role as gateway to these collections[2]. The project has been vigorously debated on discussion lists and blogs, with the participating libraries commonly referred to as "the Google 5." One point most observers seem to concede is that the questions raised by this initiative are both timely and significant.

The Google Print Library Project (GPLP)[3] has galvanized a long overdue, multifaceted discussion about library print book collections. The print book is core to library identity and practice, but in an era of zero-sum budgeting, it is almost inevitable that print book budgets will decline as budgets for serials, digital resources, and other materials expand. As libraries reallocate resources to accommodate changing patterns of user needs, print book budgets may be adversely impacted. Of course, the degree of impact will depend on a library's perceived mission. A public library may expect books to justify their shelf-space, with de-accession the consequence of minimal use. A national library, on the other hand, has a responsibility to the scholarly and cultural record and may seek to collect comprehensively within particular areas, with the attendant obligation to secure the long-term retention of its print book collections. The combination of limited budgets, changing user needs, and differences in library collection strategies underscores the need to think about a collective, or *system-wide*, print book collection—in particular, how can an inter-institutional system be

This paper originally appeared in *D-Lib Magazine*, 11,9 (September 2005).
http://www.dlib.org/dlib/september05/lavoie/09lavoie.html.

organized to achieve goals that would be difficult, and/or prohibitively expensive, for any one library to undertake individually[4]? Mass digitization programs like GPLP cast new light on these and other issues surrounding the future of library print book collections, but at this early stage, it is light that illuminates only dimly.

It will be some time before GPLP's implications for libraries and library print book collections can be fully appreciated and evaluated. But the strong interest and lively debate generated by this initiative suggest that some preliminary analysis—premature though it may be—would be useful, if only to undertake a rough mapping of the terrain over which GPLP potentially will extend. At the least, some early perspective helps shape interesting questions for the future, when the boundaries of GPLP become settled, workflows for producing and managing the digitized materials become systematized, and usage patterns within the GPLP framework begin to emerge.

This article offers some perspectives on GPLP in light of what is known about library print book collections in general, and those of the Google 5 in particular, from information in OCLC's WorldCat bibliographic database and holdings file. Questions addressed include:

Coverage: What proportion of the system-wide print book collection will GPLP potentially cover? What is the degree of holdings overlap across the print book collections of the five participating libraries?

Language: What is the distribution of languages associated with the print books held by the GPLP libraries? Which languages are predominant?

Copyright: What proportion of the GPLP libraries' print book holdings are out of copyright?

Works: How many distinct works are represented in the holdings of the GPLP libraries? How does a focus on works impact coverage and holdings overlap?

Convergence: What are the effects on coverage of using a different set of five libraries? What are the effects of adding the holdings of additional libraries to those of the GPLP libraries, and how do these effects vary by library type?

These questions certainly do not exhaust the analytical possibilities presented by GPLP. More in-depth analysis might look at Google 5 coverage in particular subject areas; it also would be interesting to see how many books covered by the GPLP have already been digitized in other contexts. However, these questions are left to future studies. The purpose here is to explore a few basic questions raised by GPLP, and in doing so, provide an empirical context for the debate that is sure to continue for some time to come. A secondary objective is to lay some groundwork for a general set of questions that could be used to explore the implications of any mass digitization initiative. A suggested list of questions is provided in the conclusion of the article.

Note on Data Sources

In the changing library landscape, the need is growing for intelligence about collections, the position of any one collection within a wider system of libraries, and important trends impacting collection management. OCLC's WorldCat bibliographic database has emerged as a strategic resource in this context: it provides the most comprehensive view available of library collections. To meet the urgent demand for more and better data, OCLC has proceeded on several fronts. It has introduced a Collection Analysis Service[5] that allows libraries to analyze and compare their collections in several dimensions. And from a research perspective, OCLC has begun looking at the characteristics of collections in systemic ways, contributing to the broad discussion that will help address issues such as those mentioned above.

The analysis that follows is based on a copy of WorldCat dating from January 2005, containing nearly 55 million records. It also uses a January 2005 copy of the WorldCat holdings file, containing nearly one billion holdings[6].

Analysis of works was based on a works index created from the January 2005 copy of WorldCat using the OCLC Research FRBR (Functional Requirements for Bibliographic Records) work-set algorithm[7].

All data and statistics reported in this article have been anonymized to avoid attaching specific data or results to specific libraries.

The System-wide Print Book Collection

Google's December 2004 press release announces its intention to "work with the libraries of Harvard, Stanford, the University of Michigan, and the University of Oxford as well as The New York Public Library to digitally scan *books* from their collections" (emphasis added). The appropriate unit of analysis for a study of GPLP, then, is a book—in particular, a print book.[8] The scope of the analysis extends to the print book collections of the Google 5, as well as to those of libraries generally.

As of January 2005, approximately one month after the Google announcement, WorldCat contained about 32 million records describing print books, or slightly less than 60 percent of the entire database. It is clear that print books account for a significant proportion of library collections, at least to the extent that these collections are reflected in WorldCat.

The 32 million books in WorldCat can be broadly interpreted as what Schonfeld and Lavoie (2005)[9] term the *system-wide print book collection*—in other words, the aggregated print book holdings across all libraries. More precisely, this total reflects the scope of the print book resource currently cataloged in WorldCat. There is a gap, of course, between these two

characterizations—the aggregate print book collection of all libraries on the one hand, and the collection of print books cataloged in WorldCat on the other. But WorldCat's status as the world's largest union catalog implies there is no other single data source representing a closer approximation to the system-wide print book collection. The 32 million print books in WorldCat, therefore, are a useful and convenient benchmark against which to consider the implications of the GPLP digitization effort; in particular, they can be viewed as an approximation of the *potential* scale of digitization that could be conducted across the system represented by the combined print book holdings of all libraries.

Coverage

The most obvious question posed by GPLP is how much of the system-wide print book collection the project would potentially cover. All discussions bearing on this issue are necessarily speculative at this point, because it has yet to be determined how much will be digitized from each library's collection. But some perspective on this issue can be obtained by looking at GPLP's *maximum coverage*—in other words, assuming each participating library's entire print book collection is digitized—and comparing this to the system-wide collection represented by the 32 million print books cataloged in WorldCat.

As of January 2005, the Google 5 have set more than 18 million holdings on WorldCat records describing print books, for an average of about 3.6 million holdings per GPLP participant.[10] This implies that the maximum potential coverage of GPLP digitization would be 57 percent of the print books cataloged in WorldCat—assuming (unrealistically) that there is no overlap at all across the print book collections of the five participating libraries.

In reality, of course, there is overlap across collections, and the degree to which it exists determines the corresponding reduction in coverage of the system-wide collection that the combined print book holdings of the Google 5 can achieve. Figure 1 illustrates actual Google 5 coverage of the system-wide print book collection, taking into account overlap across holdings for the five libraries.

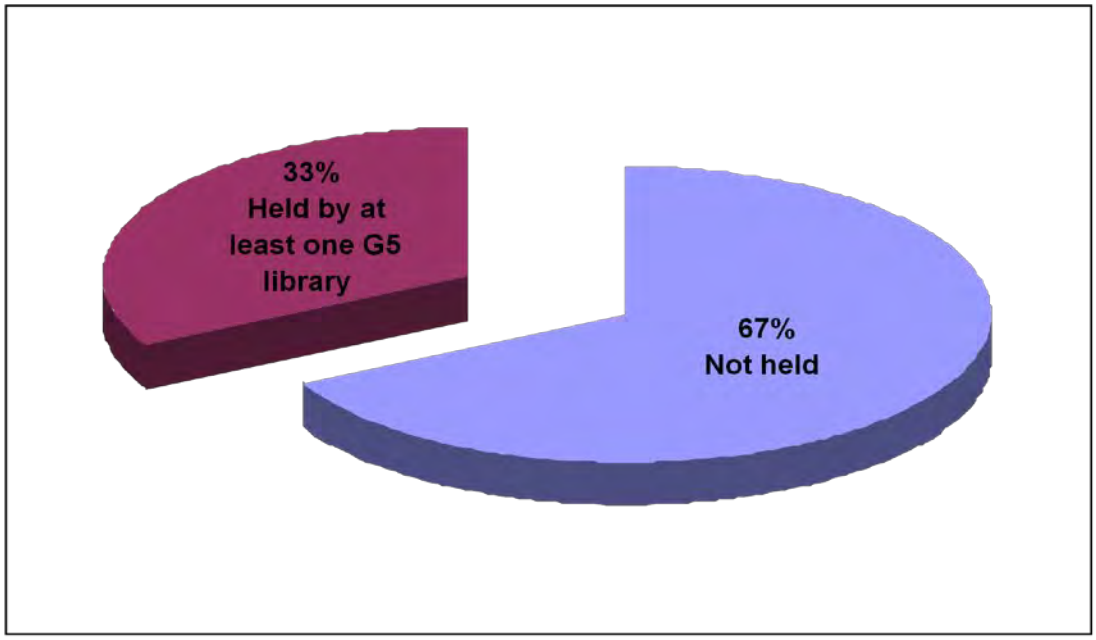


Figure 1: Google 5 Coverage of the System-Wide Print Book Collection

The proportion of the system-wide collection actually covered by GDLP, once duplicate holdings across the five institutions are removed, is about one third (33 percent), or 10.5 million unique books out of the 32 million in the system-wide collection. About two-thirds (67 percent) of the system-wide collection, or 21.6 million books, are not held by any Google 5 library.

Closer examination of the holdings data provides some insight into the degree of overlap across the Google 5 collections. Figure 2 illustrates the holdings overlap across the 10.5 million unique print books in the combined GPLP collection—i.e., the proportions held by one, two, three, four, and all five GPLP libraries.

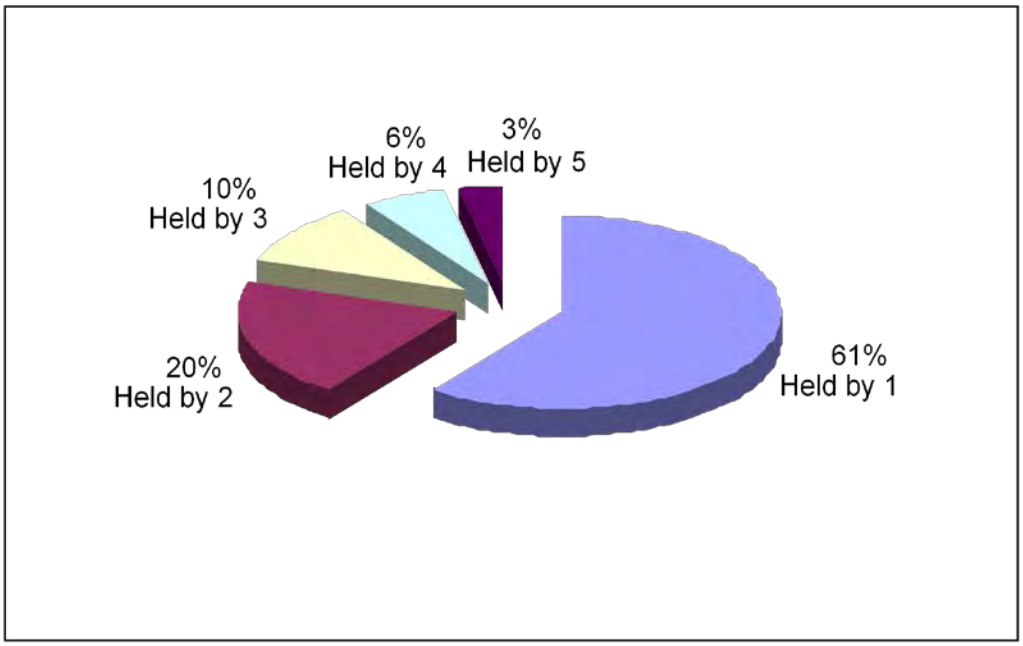


Figure 2. Google 5 Holdings Overlap

Of the 10.5 million unique books held in the combined GPLP collection, 6.3 million (61 percent) are held by only one Google 5 library; 2.1 million (20 percent) are held by two libraries; 1.1 million (10 percent) are held by three libraries; 0.6 million (6 percent) by four libraries; and 0.4 million (3 percent) by all five libraries. This pattern of cross-collection overlap implies that if each collection is fully digitized, about four out of every ten books would be re-digitized at least once, or in other words, the GPLP project reflects a minimum redundancy rate of about 40 percent.

Should this redundancy rate be considered high, low, or moderate? Several factors lead to conflicting interpretations. On the one hand, the results discussed above pertain to print book *manifestations*, where manifestation is defined according to the FRBR (Functional Requirements for Bibliographic Records) model[11]: "a physical embodiment of an expression of a work." According to this definition, two different imprints of *A Tale of Two Cities*, for example, would be considered unique books. If unique *titles* or *works* are considered, the redundancy rate may in fact be higher (see below for a more detailed discussion of this point).

However, from another perspective, overlap across the Google 5 collections can be considered quite small. The redundancy rate is, of course, likely to be a function of the number of collections being combined—the more collections, the greater the overall redundancy rate. But if analysis of overlap is confined to *bilateral* comparisons, a different picture emerges. The highest rate of print book collection overlap between two GPLP libraries is 21 percent; the lowest rate is 14 percent. The average rate is about 18 percent. This

implies that given any two Google 5 libraries—or, if the Google 5 results can be extrapolated to a larger context, given any two large research libraries—eight out of ten books in their combined collections will be unique. Of course, interpretation of this result is not straightforward, and must be considered carefully before any definitive conclusions are drawn, but at least on the surface, it does lend credence to the view that research library collections are less "vanilla" than commonly supposed.

One factor that hinders interpretation of the overall redundancy rate is that holdings overlap is often a function of the age of the book. Figure 3 illustrates the holdings overlap across the Google 5 libraries for books published in eight periods since 1800.

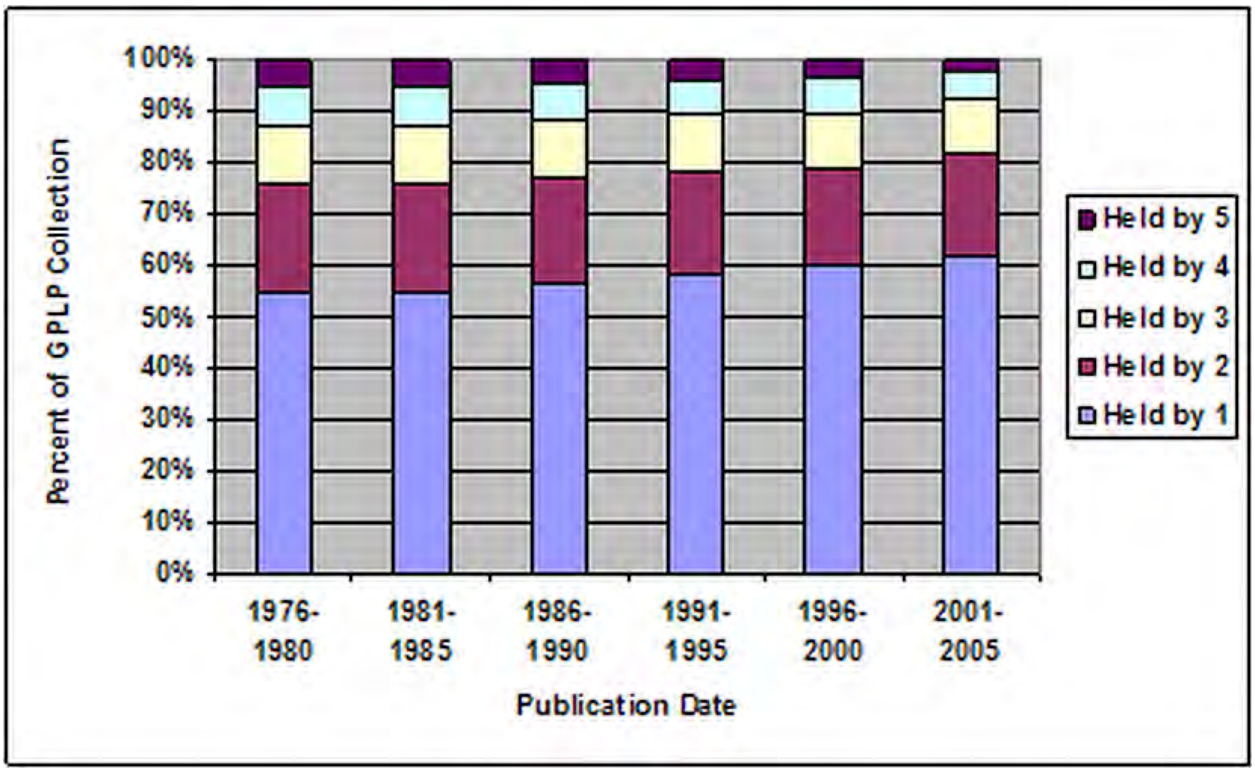


Figure 3. Google 5 Holdings Overlap, By Publication Date (1801-2005)

Figure 3 shows that the proportion of the combined GPLP collection representing uniquely held books declines as the age of the book decreases, from a high of 74 percent for books published between 1801 and 1825, to a low of 55 percent for books published between 1951 and 1975. In other words, the incidence of holdings overlap is greater for newer books compared to older ones. Interestingly, for the most recent time period (1976-2005) the proportion of uniquely held books rises slightly to 58 percent. This seemingly incongruous result warrants closer inspection.

Figure 4 offers a more granular view of holdings overlap for the period 1976–2005.

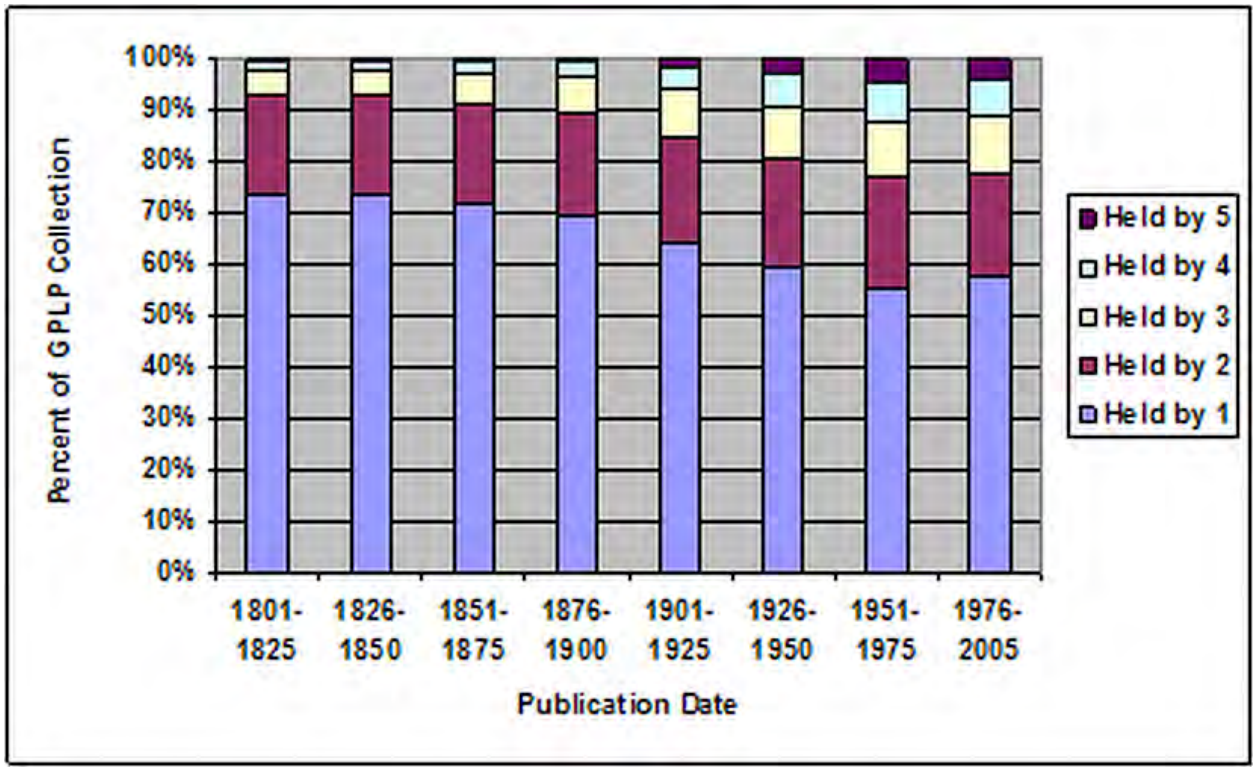


Figure 4. Google 5 Holdings Overlap, By Publication Date (1976-2005)

The proportion of books held uniquely by a single Google 5 library reaches its lowest point during the periods 1976-1980 and 1981-1985, at 55 percent. In subsequent periods, however, this proportion steadily increases—to 56 percent for 1986-1990, 58 percent for 1991-1995, 60 percent for 1996-2000, and 62 percent for 2001-2005. Lags in acquisition and cataloging are one possible explanation for this trend, although it is likely relevant only for the period 1995 to 2005. There is a possibility that these results signal a growing divergence in the collecting decisions of the Google 5 libraries in particular, and research libraries in general, but much more detailed analysis of holdings data is needed to confirm or reject this hypothesis. That is beyond the scope of this article; for the present, it must suffice to cautiously assert that the negative correlation between the age of the material in the GPLP combined collection and the degree of holdings overlap (and hence the digitization redundancy rate) seems to have reversed itself over the last twenty years.

Language

Following the GPLP announcement, there was concern in some quarters that the digitization effort would create a global resource dominated by English-language materials. These fears gained enough purchase that nineteen European national libraries recently signed an agreement to initiate a digitization program aimed exclusively at "works belonging to our continent's heritage."[12]

Some perspective on this issue can be obtained by examining the language distribution of the 10.5 million unique print books currently in the combined collection of the Google 5, as well as that for the system-wide collection as a whole.

It should be noted that WorldCat has some limitations as a data source for an analysis of this kind, since it chiefly reflects North American (and hence English-centric) library collections. Since WorldCat is used as a proxy for the system-wide collection, the latter will also exhibit a disproportionately high concentration of English-language materials, relative to the actual totality of library holdings worldwide.

Table 1 reports the distribution of languages in the combined Google 5 collection, as well as the corresponding distribution for the 32 million print books in the system-wide collection.

Table 1. Distribution of Languages: Google 5 and System-Wide Collections

Language	Google 5	System-wide
English	0.49	0.52
German	0.10	0.08
French	0.08	0.08
Spanish	0.05	0.06
Chinese	0.04	0.04
Russian	0.04	0.03
Italian	0.03	0.03
Japanese	0.02	0.04
Hebrew	0.02	0.01
Arabic	0.01	0.01
Portuguese	0.01	0.01
Polish	0.01	0.01
Dutch	0.01	0.01
Latin	0.01	0.01
Korean	0.01	0.01
Swedish	0.01	< 0.01
All others	0.07	0.08

More than 430 languages were identified in the Google 5 combined collection. English-language materials represent slightly less than half of the books in this collection; German-, French-, and Spanish-language materials account for about a quarter of the remaining books, with the rest scattered over a wide variety of languages. Corresponding results for the system-wide print book collection exhibit proportions similar to those of the Google 5 collection.

A word of explanation is useful for interpreting these results. At first glance, the fact that the combined print book holdings of four American and one British library should reflect a fifty-fifty split between English and non-English-language materials may seem incongruous. The explanation for this result lies in the effect from pooling the holdings of the five collections. The average print book collection in an English-speaking country will have a high proportion of English-language materials—perhaps on the order of 70-75 percent. But when multiple collections are pooled together, there is greater holdings overlap across English-language materials than non-English materials. Therefore, when duplicate holdings are eliminated, a larger proportion of these will be English-language materials, which in turn increases the proportion of non-English-language materials in the combined collection, relative to each individual collection. This effect will become more pronounced as more collections are added.[13]

Some corroboration for this explanation is obtained by examining the holdings overlap for English language and non-English-language print books in the combined Google 5 collection. Sixty-three percent of non-English-language print books are held uniquely by Google 5 libraries, compared to only 57 percent for English-language books. Only 6 percent of non-English language books are held by at least four Google 5 libraries, compared to 13 percent for English-language books. In short, there is a greater degree of holdings overlap for English-language print books across the Google 5 collections compared to non-English-language books, which will tend to raise the proportion of the latter in the combined collection, once duplicate holdings are removed.

It is difficult to conclude from these results whether the fears of the signatories to the European digitization agreement are justified. The combined Google 5 collection is indeed English-centric, since English-language materials account for nearly half the collection. But it is likely that many would find this proportion remarkably low.[14] Taking this into account, along with the fact that well over 400 languages are represented in the collection, suggests that the resource created by GPLP may be far more culturally diverse than originally anticipated.

Copyright

Mass digitization programs like GPLP inevitably encounter intellectual property rights issues. Indeed, on August 11, 2005, Google announced that it would temporarily suspend digitization of in-copyright books, in order to give publishers an opportunity to decide which books they would like to include—or not include—in the Google Print programs.[15] This measure, along with the intense debate over questions of copyright infringement and fair use associated with GPLP, suggests a need to examine the publication dates of the materials in the combined Google 5 print book collection.

Figure 5 shows the cumulative age distribution of the 10.5 million unique print books held by the Google 5 libraries.

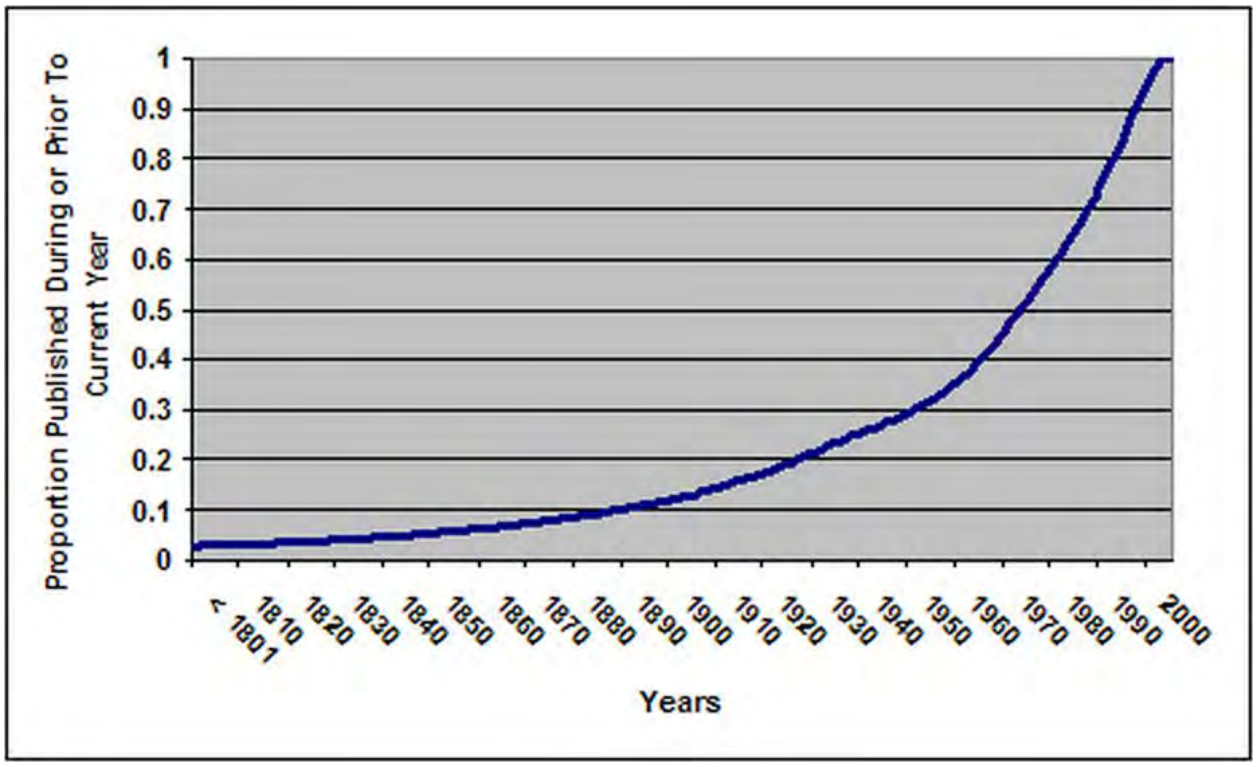


Figure 5. Cumulative Age Distribution of Google 5 Print Book Collection

Approximately half of the print books in the combined Google 5 collection were published after 1974. Almost three-quarters were published after the Second World War. Using the year 1923 as a rough break-off point between materials that are out of copyright and materials that are in copyright[16], more than 80 percent of the materials in the Google 5 collections are still in copyright.

The cumulative age distribution of the 32 million books in the system-wide print book collection is nearly identical to that of the Google 5 collection, except that the Google 5 distribution rises slightly more steeply from the early years of the twentieth century onward.

There are approximately 5.4 million books in the system-wide collection that are out of copyright. About one third of them are held by one or more of the five GPLP participating libraries. Interestingly, the Google 5 libraries hold the same proportion of the system-wide collection's in-copyright books. However, the degree of holdings overlap across the Google 5 collections for out-of-copyright print books is significantly less: more than 70 percent of out-of-copyright books are held uniquely by one GPLP library, compared to 60 percent in the overall collection.

There is some variation across the five libraries in terms of the percentage of total holdings devoted to out-of-copyright books. Three libraries each had roughly similar percentages of about 10 percent. But the other two libraries exhibited percentages nearly double that of the other three—about 18 percent. This suggests that there may be considerable differences across print book collections of large research libraries in terms of the number of out-of-copyright materials held, and by extension, the potential impact of intellectual property rights on mass digitization programs.

The proportions of out-of-copyright materials in the Google 5 and system-wide print book collections calculated based on a 1923 cut-off date should be considered a lower bound on the true values. For the years 1923 to 1963, copyright law provided that materials published during this period receive copyright protection for 28 years, which could then be renewed for an additional 47 years (now increased to 67 years according to current law). If copyright was not renewed, the material passed into the public domain.[17] If it is assumed (falsely, of course) that no materials published between 1923 and 1963 had their copyright renewed, an upper bound on the proportions of out-of-copyright materials in the Google 5 and system-wide collections can be calculated, using 1963 as the cut-off date.

Referring back to figure 5 above, and assuming all materials pre-dating 1963 are out-of-copyright, a different picture of the impact of intellectual property rights on the proposed digitization emerges. Using the 1963 benchmark date, about 63 percent of the books in the combined Google 5 collection are still in copyright, a substantially smaller proportion than that yielded when 1923 is used as the cut-off date (more than 80 percent). For the system-wide collection as a whole, the proportion is about 66 percent, compared to more than 80 percent using the 1923 cut-off date.

Looking at the approximately 10.5 million books in the system-wide collection that, according to the 1963 cut-off date, are out-of-copyright, about 36 percent are held by at least one Google 5 library, only a slightly higher proportion than that obtained when out-of-copyright is confined to pre-1923 materials only. There is greater divergence across the two copyright benchmarks, however, when considering holdings overlap for out-of-copyright print books: about 65 percent of the books are held uniquely for the pre-1963 materials, compared to about 70 percent for the pre-1923 materials (and 60 percent for the overall combined Google 5 collection).

The proportion of each library's total holdings devoted to out-of-copyright materials, where the latter is determined according to the pre-1963 benchmark, is much greater than that obtained using the 1923 benchmark, although the pattern of variation is similar. Three libraries had similar proportions of total holdings devoted to out-of-copyright books of about 28 percent. Two libraries exhibited much higher proportions: 37 and 40 percent, respectively.

Taken together, the two benchmarks dates—1923 and 1963—indicate that the proportion of the system-wide print book collection consisting of in-copyright materials, and thus potentially subject to copyright restrictions, falls somewhere between 66 and 82 percent, with the actual number dependent on the incidence of copyright renewal for materials published between 1923 and 1963. In short, at least two-thirds of the combined Google 5 collection is still protected by copyright; however, the impact of copyright restrictions on digitization of print book collections will vary across the GPLP libraries, ranging from 82 to 90 percent of holdings (according to the 1923 benchmark), or 60 to 72 percent (according to the 1963 benchmark).

Works

The FRBR bibliographic model[18] defines a work as "a distinct intellectual or artistic creation"—thus, Shakespeare's Macbeth is considered a work. An expression is "the intellectual or artistic realization of a work in the form of alpha-numeric, musical, or choreographic notation, sound, image, movement, etc., or any combination of such forms." Macbeth in the form of English-language text is an expression of the work Macbeth. Finally, a manifestation is "a physical embodiment of an expression of a work." The Folger Shakespeare Library edition of Macbeth, published in paperback by Washington Square Press in 2004, is a distinct manifestation of the work Macbeth.

In general, WorldCat records describe manifestations, and all of the results reported above pertain to manifestations. However, it is easy to imagine circumstances where digitization aimed at higher-level bibliographic entities, like expressions and works, would support the majority of potential users. Of course, there will be some cases where digitization of specific imprints or even specific copies will be important to some users, but the cost of supporting these users may be prohibitive. In this case, the goal of a digitization initiative may be to digitize a single exemplar manifestation, rather than multiple manifestations, of a work or expression.[19]

OCLC Research has developed an algorithm[20] that converts MARC21 bibliographic databases into FRBR work sets, where a work set is a cluster of WorldCat records—i.e., manifestations—pertaining to the same work. This algorithm was applied to the January 2005 copy of WorldCat used in this study in order to obtain some perspective on the implications of GPLP in terms of works.

The 32 million manifestations in the system-wide print book collection can be rolled up into approximately 26.1 million distinct works. Each of these works contains an average of only 1.2 print book manifestations—essentially, one print book manifestation per work. Note that for the purposes of this analysis, only manifestations in the form of print books are considered; other manifestations, such as those in digital or audio formats, are excluded from the analysis.

Total holdings set by the Google 5 libraries on the 26.1 million works containing at least one print book manifestation are about 16.7 million (note that all holdings set by a single library for multiple manifestations of the same work are counted as one holding). Figure 6 illustrates Google 5 coverage of manifestations and works.

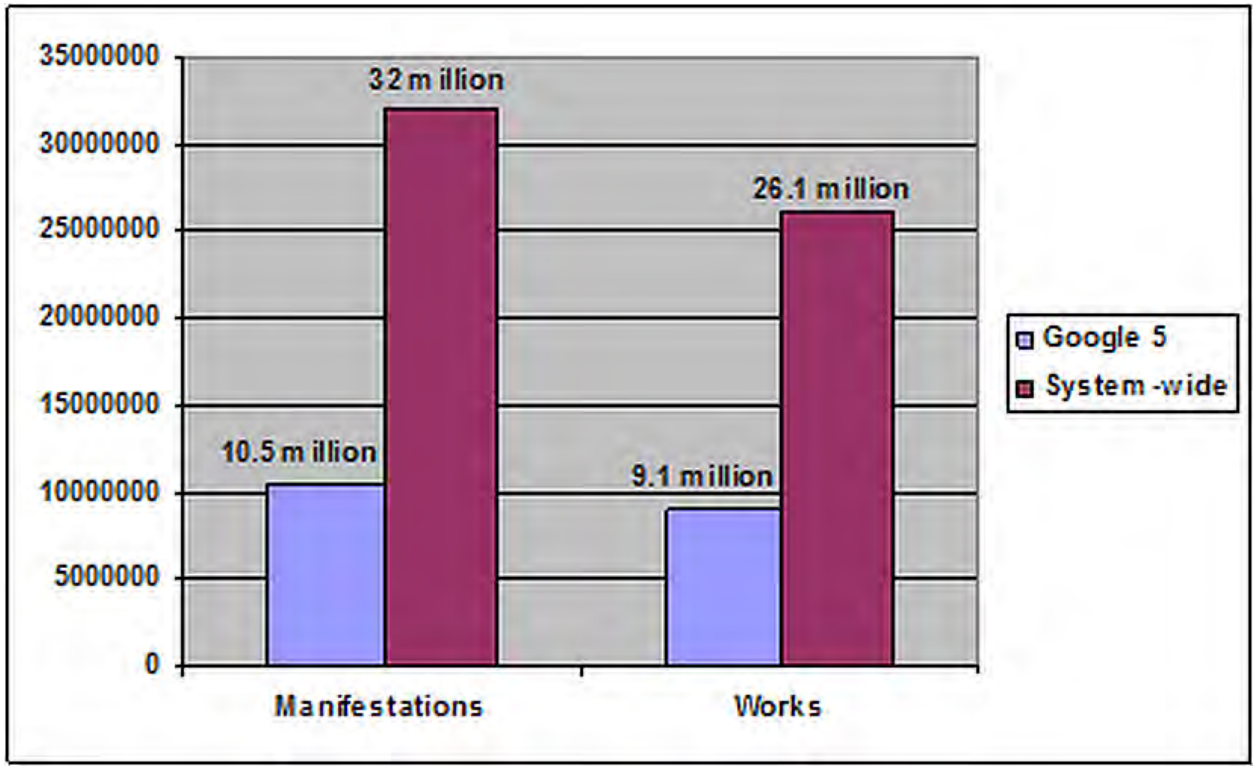


Figure 6. Google 5 Coverage: Manifestations and Works

Of the 26.1 million distinct print book works, about 9.1 million, or 35 percent, are held by at least one GPLP library, indicating that GPLP coverage in terms of works is only slightly higher than in terms of manifestations.

About 56 percent of works are held uniquely by one Google 5 library, compared to about 60 percent for manifestations. This result accords with intuition, since aggregating manifestations into works should reduce the overall "uniqueness" of the collection. However, this reduction is only slight, most likely because the majority of works have one, or at most only a few, manifestations. At the other end of the holdings distribution, about 12 percent of works are held by at least four Google 5 libraries, compared to 9 percent for manifestations.

Forty-four percent of the works are held by two or more Google 5 libraries, which suggests that digitization of the full print book collections of the Google 5 would result in a little more than four out of every ten digitized books being redundant, assuming digitization of works (or titles), rather than manifestations, was the goal of the project. This is virtually the same redundancy factor estimated for digitization of manifestations, again due to the

fact that most works have only a few manifestations. But this result masks the fact that there is likely to be a "core" set of widely held works, each with many manifestations, for which the redundancy rate will be extremely high. For this core set of works, there may be significant scope for cost savings if digitization focuses on works or expressions, rather than manifestations.

Convergence

Those who see positive implications for GPLP may count among its merits the possibility that it will serve as a first step toward the larger goal of digitizing and making available online the full print book collections of libraries all over the world. However, achieving this goal will not be easy. Recent work by Schonfeld and Lavoie (2005)[21] suggests that the system-wide print book collection (as reflected in WorldCat) is dispersed widely over many institutions. Nearly 40 percent of all print books are held uniquely by one institution. Only a third of print books have more than 5 holdings; about half have two or fewer holdings. This suggests that the system-wide print book collection is dispersed over many institutions, and that many books are "rare," in the sense of not being widely held. There is a need for further work to ascertain the characteristics of these rare materials, and determine their importance to mass digitization efforts.

As noted above, the GPLP stands to cover approximately one third of the system-wide print book collection. Attaining this degree of coverage by aggregating the holdings of only five large libraries is a remarkable achievement, but it also poses two questions: first, what would be the results if a different set of five libraries had participated in GPLP? And second, what incremental extensions to coverage can be obtained by adding additional libraries to the original Google 5?

To provide some very rudimentary perspective on these questions, five additional libraries were selected (in no particularly systematic way) to include in the analysis: a small American liberal arts college, a large Canadian university, a large American public university, a large American private university, and a large American metropolitan public library[22]. This selection is as US-centric as the original Google 5, but in a sense this is appropriate, given that WorldCat largely reflects North American library collections. Holdings data can be used to assess the impact on coverage of the five collections in aggregate, as well as each individual collection.

Taken together, the five new collections account for approximately 8 million holdings, compared to more than 18 million for the original Google 5. The disparity in total holdings is largely because the new collections exhibited more variance in size: in the original Google 5, the largest collection was a little more than double the size of the smallest; in the five new collections, the largest collection is almost nine times the size of the smallest.

The combined holdings of the five new libraries account for about 5.9 million unique print books, or 18 percent of the system-wide collection of 32 million books. This is much less than the 10.5 million books from the original Google 5, but if the results are weighted to adjust for the disparity in number of holdings between the Google 5 collection and the new collection, a different picture emerges. Computing the ratio of unique print books to total holdings for each combined collection yields 74 percent for the new collection, compared to only 58 percent for the Google 5 collection. This indicates that the degree of redundancy associated with the new collection is less: digitization of four out of every ten Google 5 books would be redundant; only 2 to 3 books out of every ten for the new collection would be redundant.

A smaller degree of redundancy for the new collection is also suggested by an examination of the distribution of holdings across the five new libraries. Of the 5.9 million unique books in this collection, nearly three quarters are held uniquely by a single library, compared to only 60 percent for the Google 5. About 9 percent of the Google 5 print books were held by at least four Google 5 libraries; only about 1 percent of the books in the new collection are held by at least four libraries.

Bilateral comparisons between the combined Google 5 collection and each of the five new collections yield insight on the impact on coverage obtained by adding the print book holdings of various library profiles. In absolute terms, the large American private university added the greatest number of unique books—about 1 million—to the existing Google 5 total, a 10 percent increase. The small American liberal arts college added the fewest unique books—about 71,000—for an increase of less than 1 percent. The large American public university was second with nearly half a million books (5 percent increase); the large American metropolitan public library was third with a little more than 231,000 books (2 percent increase); the large Canadian university was fourth with about 104,000 books (1 percent increase).

These results are partly a consequence of the disparity in collection sizes, as reflected in WorldCat holdings: the large American private university had the most holdings of the five, and the small American liberal arts college the second least. A rough way to adjust for collection size is to compute the ratio of unique books added to the Google 5 collection as a percent of the institution's total holdings. From this perspective, the large American metropolitan public library exhibited the highest degree of uniqueness relative to the Google 5 collection: 39 percent of its holdings were unique relative to the combined Google 5 holdings. The large American private university was next at 25 percent, followed by the large Canadian university (23 percent), the large American public university (21 percent), and the small American liberal arts college (13 percent).

Finally, the combined collections of the original Google 5 on the one hand, and the five new libraries on the other, were compared. The two combined collections together account for about 12.3 million books, an increase of about 1.8 million books, or about 17 percent, over

the Google 5 collection alone. This result suggests that digitization of the full system-wide print book collection will require the participation of many libraries of all types: adding nearly 8 million new holdings from a variety of library types to those of the Google 5 collection was sufficient to account for only 8 percent of the print books not held by one or more of the Google 5 libraries. It is likely that if a second new collection of five libraries were added to this total, the returns, measured in additional unique books, would be smaller still.

Conclusion

If it ends up proceeding along the lines of its original plan, the Google Print Library Project promises to be significant both for libraries and their users—but it is still early days, so the precise nature of that significance is yet to be discerned. Even if it does not, GPLP at the very least offers an interesting test case with which to think about the implications of multi-institution mass digitization programs. Speculation on what directions GPLP will take in the future, and the resultant impact on libraries, will, of course, continue. This article suggests a number of areas where an impact will likely be felt—coverage, language, copyright, works, and convergence—and supplies some empirical context for thinking about issues related to these areas.

GPLP is only one of what will likely be many mass digitization programs underway in the near future. As these projects emerge, it would be useful to have at hand a set of general questions with which to consider their implications for libraries and users. The analysis reported in this article motivates a starter list of questions useful for considering the implications of multi-institution mass digitization programs:

- What are the characteristics of the overarching "population" of materials that will serve as the target of the digitization effort (e.g., the system-wide print book collection)?
- How much of this population will the digitization effort potentially cover?
- What is the degree of redundancy associated with the digitization effort?
- What bibliographic unit is the focus of digitization (e.g., manifestations, expressions, works)?
- What number of participants and combination of institution types is optimal for obtaining the maximum benefit with the minimum cost, in relation to achieving a particular set of digitization goals?

As mass digitization programs become more common, many are likely to originate within the library community itself, rather than through external organizations like Google. For library-

initiated (and funded) programs especially, it is imperative that digitization efforts 1) are organized in ways that leverage available resources to maximize community benefits, and 2) reflect a digitization strategy that is conscious of system-wide implications. Careful analysis of proposed digitization programs, using the best data sources at hand, helps decision-makers anticipate and shape the impact of these programs in ways that contribute toward the realization of both of these objectives.

Acknowledgments

The authors would like to thank Dale Flecker, Clifford Lynch, Ed O'Neill, Donald Waters, and John Price Wilkin for reading and commenting on an earlier draft of this article.

Notes and References

1. See http://www.google.com/press/pressrel/print_library.html. It should be noted that on August 11, 2005, Google announced a temporary suspension of digitization of in-copyright books, in order to give publishers an opportunity to decide which books they would like to include in (or exclude from) the Google Print program.
2. For an overview of various perspectives on the Google Print Library Project, see Roush, Wade. 2005. "The Infinite Library: Does Google's Plan to Digitize Millions of Print Books Spell the Death of Libraries; or Their Rebirth?" *Technology Review.* (1 May).
3. See http://print.google.com/googleprint/library.html for a description of the project.
4. For example, there is discussion about backup depositories, including their coordination and shared attention to withdrawal of books. More generally, in a network environment users are becoming used to interacting with resources without regard to location, and most libraries provide only a part of the collection that might be of use.
5. See http://www.oclc.org/collectionanalysis/ for more information about this service.
6. Note that multiple copies of the same book count as only one holding.
7. See http://www.oclc.org/research/projects/frbr/algorithm.htm.
8. Although there is no unambiguous bibliographic definition of a book, libraries have often used monographic language materials as a proxy for books, and this practice is adopted for this study. More specifically, in the context of a MARC21 record, a book is defined as a language-based monograph, identified by the codes "a" and "m" in bytes 6 and 7 of the leader, respectively. For the purposes of this study, theses/dissertations and government documents are excluded from the analysis, since these materials are usually acquired and managed as separate segments of the library collection. Records describing books in print format were identified by eliminating all non-print formats, such as digital, microform, Braille, and so on.
9. Schonfeld, Roger and Brian Lavoie. 2005. "Characterizing the System-Wide Collection" (paper in preparation). Preliminary findings were reported in "A System-Wide View of Library Collections." Presented at the Spring 2005 CNI Task Force Meeting. http://www.oclc.org/research/presentations/lavoie/cni2005.ppt.
10. Note that the 18 million holdings reported here reflect the fact that duplicate holdings across library units within the same institution have been removed.
11. See http://www.ifla.org/VII/s13/frbr/frbr.pdf.
12. See http://www.dw-world.de/dw/article/0,1564,1566717,00.html for a description of this initiative.

13. For example, suppose there are two library collections, each consisting of 10 books, 7 of which are English, and 3 non-English. So each collection has a 70-30 split between English- and non-English-language books. Now suppose that 5 out of the 14 total English-language book holdings, and 1 of the 6 total non-English-language book holdings, are duplicates. Combining the two collections and eliminating duplicate holdings results in 14 unique books, 9 of which are English and 5 of which are non-English, for a 64-36 split in the combined collection
14. Moreover, it should be noted that some of the English-language books will be translations into English from other languages.
15. See http://googleblog.blogspot.com/2005/08/making-books-easier-to-find.html.
16. The use of 1923 as a break-off point is in reference to US copyright law. Of course, materials published outside the US are not necessarily subject to US copyright laws, but the US copyright regime was chosen as the benchmark to simplify the analysis. This analysis could be repeated for other copyright regimes.
17. According to current US copyright law, materials published in the period 1963-1977 receive copyright protection for 28 years, plus an automatic extension of 67 years; therefore, these materials should still be in copyright, as well as all materials published after 1977. See http://www.cepic.org/html/budapest/lawusa.htm for a brief overview of past and present US copyright regimes.
18. See http://www.ifla.org/VII/s13/frbr/frbr.pdf.
19. More precisely, digitization would probably focus on expressions, rather than works. An English-language textual version and a French-language textual version are both distinct expressions of the work Macbeth, but it is unlikely they would be considered substitutes, in the sense that digitizing one would eliminate the need to digitize the other. However, there is still much debate over how to identify expressions in bibliographic records, and for this reason, the remainder of this section focuses on works.
20. See http://www.oclc.org/research/projects/frbr/algorithm.htm.
21. See note 9 above.
22. The Carnegie classification for the small American liberal arts college is “Baccalaureate Colleges—Liberal Arts.” The Carnegie classification for the large American public and private universities is “Doctoral/Research Universities—Extensive.” The large Canadian university and large American metropolitan public library are not included in the Carnegie classifications.

Beyond 1923:
Characteristics of Potentially In-copyright Print Books in Library Collections

Brian Lavoie and Lorcan Dempsey

Introduction

Issues of copyright and permissible use have swirled around efforts to digitize print book collections. Sharp debate has ensued over the circumstances in which creating a digital surrogate and making it accessible online runs afoul of copyright protections, and what remedies might be appropriate to compensate rights holders. Some digitization efforts, such as the Open Content Alliance, have restricted themselves to public domain materials; Google Books, on the other hand, has sought to reach agreement with copyright holders represented by the Authors Guild and the Association of American Publishers. A proposed class-action settlement,[1] announced in October 2008, would create a Book Rights Registry responsible for administering and adjudicating the process of locating and compensating rights holders impacted by Google's digitization activities.

The Google book settlement provoked spirited discussion of its potential ramifications, mimicking the commotion that followed the announcement of the original Google Print for Libraries (later re-named Google Books) project in December 2004. Using data from the WorldCat bibliographic database,[2] OCLC Research published an article in 2005 aimed at illuminating issues surrounding Google's plan to digitize the print book collections of five major research libraries. The present article is motivated by a similar purpose: to provide empirical context for the many discussions surrounding the digitization of in-copyright print books. The settlement has raised challenging questions regarding permissible use of print book titles published after 1923; many of these titles may eventually form a significant part of the Google book database should it come to pass.

This paper originally appeared in *D-Lib Magazine*, 15,11/12 (November/December 2009). http://www.dlib.org/dlib/november09/lavoie/11lavoie.html.

Discussions of Google Books and other digitization efforts tend to treat in-copyright print books as an amorphous collection, with little elaboration or detail on what this important collection of materials actually looks like. How many titles are involved? What is the distribution of their publication dates? What general observations can be made about their content? This article examines these and other questions in regard to the collection of US-published print books represented in WorldCat. Many of these questions were posed to the authors in private inquiries; these inquiries, along with the keen interest in digitization that continues to spark debate on blogs and listservs, suggested that a general publication addressing the characteristics of in-copyright print books could provide helpful context for ongoing discussions.

The focus of this article is on print book titles that are either in-copyright or potentially in-copyright. Determining copyright status is, however, problematic. The nuances of US copyright law are quite complicated, but a useful simplification organizes print books into three categories of copyright status based on date of publication. Broadly speaking, works published before 1923 are considered in the public domain, and therefore unencumbered by copyright restrictions. The copyright status of books published between 1923 and 1963, however, is murkier. Under US copyright law, works published during this period with a copyright notice remain in copyright for 95 years after publication—*if their copyright was renewed*. If copyright was allowed to lapse, the work reverts to the public domain. Finally, books published after 1963 are, by and large, still in copyright.

In addition to copyright status, the question of *orphan works* has received much attention in regard to digitization activities. The United States Copyright Office defines an orphan work as "the situation where the owner of a copyrighted work cannot be identified and located by someone who wishes to make use of the work in a manner that requires permission of the copyright owner."[3] While it is important to bear in mind that any in-copyright book can be an "orphan," in practice the prevalence of orphan works is likely to be skewed toward older, rather than recently published, materials.

The analysis that follows examines the characteristics of US-published print books, with an emphasis on books that are likely in copyright according to US copyright law.[4] As with our earlier article, the analysis is based on data from the WorldCat database, which represents the aggregated collections of more than 70,000 libraries worldwide. The analysis focuses on three areas: the WorldCat aggregate collection of US-published print books; the subset of this collection published during or after 1923—i.e., those potentially associated with copyright and/or orphan works issues; and the combined print book collection of three academic research library participants in Google Books—again, with an emphasis on materials that are potentially in copyright.

Characteristics of the Aggregate US-published Print Book Collection in WorldCat

As of April 2009, the WorldCat bibliographic database contained about 135.3 million bibliographic records representing information resources of all descriptions. Of these, 104.1 million represented books, and of these, 84.8 million were print books.[5] Finally, of these, 15.5 million were print books published in the US, and therefore presumably covered by US copyright law (figure 1). It is important to keep in mind that these counts do not represent "physical objects"—i.e., copies of books on the shelf—but rather, distinct imprints or *manifestations*.[6] For example, the Short Books, Limited publication of the book *Walking Ollie*, published in London in 2006, is a distinct print book manifestation; the version of the same book published by Perigee Books (New York, 2008) is another distinct manifestation. There are likely hundreds if not thousands of physical copies of these two manifestations worldwide in the thousands of institutional print book collections represented in WorldCat. However, each manifestation would only be counted once in WorldCat.

For the remainder of this article, we will refer to these 15.5 million print book manifestations as the aggregate US-published print book collection in WorldCat.

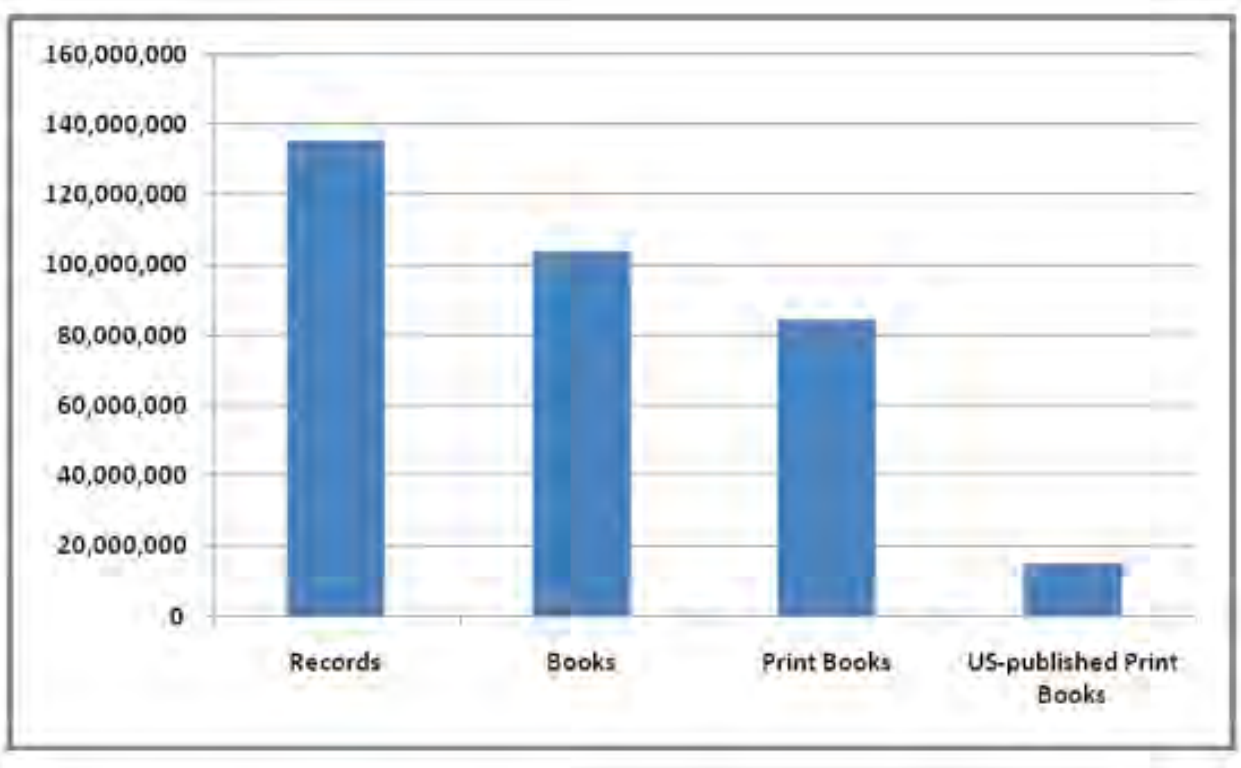


Figure 1. US-published Print Books in WorldCat: April 2009

The aggregate US-published print book collection is a resource that is collectively managed by many institutions; the 15.5 million distinct manifestations in the aggregate collection inflate to 656.8 million total holdings in library collections around the world, where a "holding" simply means that a particular library collection contains at least one copy of a particular print book manifestation. In fact, the 15.5 million US-published print book manifestations—constituting only 11 percent of the materials represented in the WorldCat database—account for 46 percent, or nearly half, of the more than 1.4 billion total holdings attached to materials represented in WorldCat. To some extent, this result is predictable: North American library collections are especially well-represented in WorldCat, and one would expect North American libraries to collect US publishing output heavily. Certainly these libraries also collect much of the rest of the world's published output as well, but their combined collections yield a strong concentration of US-published print books.

Probing a little deeper into the total holdings of the aggregate US-published print book collection reveals that a wide range of institution types is represented within the group of libraries that collectively hold this resource. Table 1 lists the percentage of total holdings attributable to different institution types for the aggregate US-published print book collection.

Table 1. Total holdings of the aggregate US-published print book collection, by institution type

Period	Number	Percentage
Academic	363,542,724	56 percent
Public	219,920,744	33 percent
Special	20,898,191	3 percent
School	19,147,728	3 percent
Other Government	11,816,402	2 percent
State & National	8,660,213	1 percent
Other	5,816,423	1 percent
Type unknown	7,021,683	1 percent

More than half of the holdings attached to the 15.5 million US-published print book manifestations belong to academic institutions; while a third belong to public libraries, and the rest to a variety of other institution types. This suggests that among collecting institutions, academic libraries possess an especially considerable stake in issues impacting accessibility, use, and preservation of US-published print books, if for no other reason than by virtue of the comparatively large investment they have made in collecting them, and the consequently large presence these materials have in their collections. The age (i.e., publication date) of the titles in the US-published aggregate print book collection is not

distributed evenly over time, but instead is skewed toward newer materials. Figure 2 shows the distribution of US-published print book manifestations in WorldCat, by publication year, for the period 1900-2008.

Figure 2. US-published print book manifestations, by publication date (1900-2008)*

** Note: the drop in manifestation totals after 2000 is not a consequence of decreased publishing output or collecting activity in those years, but instead represents "cataloging lag," the elapsed time between a book's date of publication and the date that a bibliographic record for the book is entered into the WorldCat database. Cataloging lag can be divided into two components: 1) the acquisition lag, which is the time elapsed between the date of publication and the time it is acquired by a library; and 2) the processing lag, which is the time between the date of acquisition and the date a record for the book is entered into WorldCat.*

Approximately half of the manifestations in the aggregate US-published print book collection in WorldCat were published after 1977; two-thirds were published after 1964, and three-quarters after 1951. This would suggest that the number of US-published print book manifestations in library collections that are "free and clear" in terms of copyright restrictions is comparatively small, while the fraction that is likely in copyright is comparatively large. Table 2 sharpens this point by organizing US-published print book manifestations in WorldCat by the three broad time frames relevant to assessing copyright status.

Table 2. Distribution of US-published print book manifestations in WorldCat, by major US copyright period

Period	Number	Percentage
Pre-1923	2,227,048	14 percent
1923-1963	2,596,114	17 percent
Post-1963	9,991,301	65 percent
Unknown/questionable date	667,814	4 percent

The percentages reported in Table 2 indicate that about 14 percent of the US-published aggregate print book collection was published before 1923, and therefore is, with reasonable certainty, in the public domain according to US copyright law. A further 17 percent were published between 1923 and 1963; for these, copyright status cannot be ascertained without investigating each individual title. Some portion of these materials will be in the public domain—in particular, those whose copyright was not renewed. The rest will still be under copyright. Recent statistics from the HathiTrust indicate that about 60 percent of candidate materials for digitization published between 1923 and 1963 reverted to the public domain, either because copyright was not renewed; the book was published without a copyright notice, or for other reasons.[7] Applying this fraction to the US-published aggregate print book collection in WorldCat suggests that approximately 1.6 million manifestations are public domain, while the remaining 1 million are still in copyright.

The HathiTrust result is based on academic library holdings, while the aggregate print book collection in WorldCat represents the holdings of a variety of institution types (although as Table 1 indicates, academic libraries hold the largest portion). A more general, but much earlier study by the US Copyright Office in 1960 found that only 7 percent of books registered for copyright in 1931-32 had had their copyright renewed within the prescribed 28-year period after initial registration. The remainder of the books would have reverted to the public domain.[8] Both the HathiTrust and Copyright Office results suggest that of the print books published between 1923 and 1963, a majority—and perhaps a substantial majority—are likely to be in the public domain.

Finally, about two-thirds of the US-published aggregate print book collection represents books that were published after 1963, and therefore are almost certainly still in copyright.

As noted at the beginning of this article, an orphan work is any work under copyright where the rights holder cannot be identified or located. In light of the US-published aggregate print book collection in WorldCat, this issue is relevant for the 12.6 million print book manifestations that could *potentially* still be in copyright—the sum of the totals for the two periods 1923-1963 and post-1963 in Table 2, or 82 percent of the collection. Of course, in

practice the fraction that end up as orphan works will be considerably smaller; the 12.6 million manifestations is appropriately interpreted as the pool of materials from which orphan works could potentially emerge. However, even if but a fraction of these manifestations end up as orphan works—say as little as 10 percent—that would still account for over 1 million manifestations where administration of copyright permissions is inhibited by a lack of a clear rights holder. And this is not the full extent of the orphan works problem: orphan works can emerge from books published outside the US as well.

Characteristics of Potentially "In Copyright" US-published Print Books

In this section, the characteristics of the 12.6 million print book manifestations published during or after 1923—that is, the pool of print book manifestations in library collections that are potentially in copyright—are examined as an aggregate collection. It is these materials that pose potential copyright-related difficulties in the context of print book digitalization efforts such as Google Books and others. A number of aspects of these materials are examined in this section, including authors, subject, and audience level, with a special focus on nonfiction materials.

The aggregate collection of potentially in-copyright, US-published, print book manifestations in WorldCat is associated with about 3.7 million unique authors (individuals)[9] who served some creative role in regard to one or more of the print book manifestations in the collection, and therefore could potentially hold some form of copyright authority over future use of the book. Table 3 reports the authors associated with the most print book manifestations in WorldCat that are potentially still in copyright.

Table 3. Authors associated with the most potentially in-copyright print book manifestations

Author	Number of Print Book Manifestations
William Shakespeare	6,226
Carole Marsh	3,295
Mark Twain	2,979
Jack Rudman	2,554
Charles Dickens	2,359
Ronald Vern Jackson	2,265
Harold Bloom	2,234
Agatha Christie	2,061
Robert Louis Stevenson	1,946
Joy Cowley	1,877

Note that this list is not intended to imply that these are the most "important" authors (although some clearly are!); only that these are the most frequently appearing authors in the collection under study. A perhaps surprising feature of this list is that many of the authors are those whose work one might suppose had long since passed into the public domain. Authors whose works have become time-honored classics will be published and republished frequently over time; intellectual property rights will likely attach to the various manifestations of these works, in the form of new introductions, illustrations, and even updates, revisions, or new representations of the "main body" of the work itself that constitute sufficient grounds to claim "new material." The names on the list in Table 3 serve as a useful reminder that digitization efforts may encounter copyright issues even with works originally published centuries ago; recently published manifestations of Shakespeare's *Macbeth*, or Dickens' *A Christmas Carol*, for example, are likely not in the public domain.

Of the 12.6 million print book manifestations published during or after 1923, the overwhelming majority—92 percent, or 11.6 million—were nonfiction; only 8 percent, or about 1 million, were fiction. For the remainder of this section, statistics reported pertain to nonfiction print book manifestations only.

Table 4 presents a breakdown of the post-1923 nonfiction print book manifestations by subject.[10]

Table 4. Subject breakdown, nonfiction print books

History and auxiliary sciences	8 percent
Engineering and technology	7 percent
Business and economics	7 percent
Language, linguistics, and literature	6 percent
Philosophy and religion	5 percent
Health and medicine	5 percent
Art and architecture	3 percent
Law	3 percent
Sociology	3 percent
Education	3 percent
Other	15 percent
Unknown	35 percent

The methods used in this study were able to assign subject categories to about two-thirds of the nonfiction print book manifestations. The distribution of the categorized manifestations over subject was fairly even, with no category exceeding 8 percent of the total. History is the most populous category, followed by Engineering & Technology and Business & Economics. These categories account for nearly a quarter of the nonfiction manifestations.

In addition to subject, it is also possible to make some observations about the "audience level" of the nonfiction print book manifestations. Audience level is an inference about the nature of the content of a book. While audience level cannot be determined directly, it is possible to make some useful inferences based on the collecting decisions of libraries. For example, if a particular book is held primarily by academic libraries, we can infer that it is likely intended for a scholarly or research-oriented audience. If a book is held chiefly by primary or secondary education institutions, it is likely intended for a juvenile audience. Examining the library holdings attached to a particular print book manifestation in WorldCat, it is possible to calculate a metric whose value ranges from zero to one, where values closer to zero indicate the book is held mainly by primary or secondary education institutions, and values closer to one would indicate the item is held mainly by academic libraries.[11] For simplicity, we divide this continuum up into three categories:

Audience level from 0 to 0.33: "Schooler" (primarily intended for a juvenile audience)

Audience level from 0.33 to 0.67: "General" (primarily intended for a nonspecialist readership)

Audience level from 0.67 to 1.0: "Scholar" (primarily intended for an academic or specialist readership)

It should be noted that the audience level metric is only an estimate, and may not be accurate for any particular book. However, as a means of making some general observations about the broad contours of an aggregate collection of books, it does provide some useful insights.

Application of the audience level metric to the collection of US-published nonfiction print books published during or after 1923 is limited to books whose records were entered into WorldCat prior to January 2007.[12] This restriction helps ensure that the book's record has resided in WorldCat long enough to accumulate a sufficiently representative collection of holdings. This reduces the pool of print book titles to about 8 million. Table 5 reports the audience level calculations for these materials.

Table 5. Audience level breakdown, US-published nonfiction print books published during or after 1923*

Schooler	4 percent
General	42 percent
Scholar	54 percent

** For manifestations where it was possible to calculate an audience level (92 percent of total). Most of the manifestations without an audience level were materials held by a type of library not considered in the audience level calculations.*

The vast majority of the nonfiction print book manifestations (96 percent) were intended for general readership or higher; the fraction intended for juvenile audiences was quite small (4 percent). The fraction devoted to a scholarly or specialist audience constituted the largest fraction of the titles. This finding complements the results in Table 1, which indicate that the majority of US-published print book holdings in library collections are attributable to academic institutions.

More nuanced results are obtained by examining audience level calculations for several of the subject categories identified in Table 4. Table 6 reports the audience level breakdown for the top six subject categories in Table 4. As with Table 5, the results reported in Table 6 are limited to print books entered into WorldCat prior to 2007.

Table 6. Audience level breakdown, US-published nonfiction print books published during or after 1923: by subject*

Subject	Schooler	General	Scholar
History	6 percent	58 percent	36 percent
Engineering/Technology	3 percent	45 percent	51 percent
Business/Economics	1 percent	34 percent	66 percent
Language/Linguistics/Literature	8 percent	40 percent	52 percent
Philosophy/Religion	3 percent	52 percent	45 percent
Health/Medicine	2 percent	31 percent	67 percent

** For manifestations where it was possible to calculate an audience level. Fraction of manifestations with an audience level in each subject category: History (94 percent); Engineering/Technology (90 percent); Business/Economics (92 percent); Language/Linguistics/Literature (99 percent); Philosophy/Religion (97 percent); Health/Medicine (95 percent).*

According to the results in Table 6, all subject categories had relatively small fractions of materials intended primarily for juvenile audiences; Language/Linguistics/Literature exhibited the highest fraction with 8 percent. History was the subject category with the highest proportion of materials aimed at a general or nonspecialist audience (58 percent); Philosophy/Religion was the only other category where the proportion of General materials exceeded the proportion of Scholar materials. Health/Medicine was the most "scholarly" subject category, with two-thirds of print books aimed at a research or specialist audience;

Business/Economics was slightly behind Health/Medicine at 66 percent. Taken together, the results in Table 6 indicate that digitization efforts aimed at subject categories like Health/Medicine and Business/Economics would most likely confer the most benefits on a scholarly or specialist audience; in contrast, digitization aimed at History or Philosophy/Religion would perhaps be of greater benefit to nonspecialist readers.

The preceding analysis has focused on the accumulated US nonfiction publishing output from 1923 to the present. To gain a different perspective on the characteristics of these materials, print books published in 1923 and print books published in 2000 were identified and compared as a means of examining changes in the characteristics of nonfiction print books in WorldCat over time. About 40,000 US-published nonfiction print book manifestations published in 1923 reside in WorldCat, compared to about 235,000 published in 2000. Table 7 reports the subject category breakdown for these two collections of print books.

Table 7. Subject breakdown, US-published nonfiction print books (1923 and 2000)

1923	
Language, Linguistics and Literature	10 percent
History and Auxiliary Sciences	9 percent
Philosophy and Religion	6 percent
Business and Economics	5 percent
Engineering and Technology	4 percent
Health and Medicine	3 percent
Other	18 percent
Unknown	44 percent
2000	
History and Auxiliary Sciences	11 percent
Business and Economics	7 percent
Engineering and Technology	7 percent
Language, Linguistics and Literature	6 percent
Philosophy and Religion	6 percent
Health and Medicine	6 percent
Law	4 percent
Art & Architecture:	4 percent
Sociology	4 percent
Education	4 percent
Computer Science	3 percent
Other	14 percent
Unknown	24 percent

Key differences between print books published in 1923 and those published in 2000 include a much larger proportion devoted to Language/Linguistics/Literature in the earlier period, while a larger proportion was captured by Engineering/Technology and Business/Economics in the later period. And of course, Computer Science as a subject category only makes an appearance in the latter year. The results in Table 7 suggest that the so-called STM (science, technology, medicine) subject categories account for a significantly higher proportion of the print book manifestations in 2000 than in 1923. For example, Engineering/Technology and Health/Medicine together account for 13 percent of the manifestations in 2000, compared to 7 percent in 1923. In contrast, the humanities are slightly more predominant in 1923 compared to 2000: History, Language/Linguistics/Literature, and Philosophy/Religion account for 25 percent of the titles in 1923, compared to 23 percent in 2000. However, these findings should be considered preliminary results; more work needs to be done to verify and sharpen their implications. A large percentage of print books published in 1923 were not categorizable by subject according to the methods used in this study. Allocation of these uncategorized manifestations to the various subject categories could shift the proportions significantly. This consideration also impacts the results for 2000, but to a lesser extent: in this case, only about a quarter were not categorized.

Another factor impacting interpretation of the findings in Table 7 is the question of whether the distribution of manifestations across subject categories is influenced more by the scope of publishing output in each year, or by the collecting and retention decisions of libraries. It is not possible to resolve this question from the data in Table 7, but it is an important issue in understanding the dynamic character of the collection of US-published print books in WorldCat.

Table 8 reports the audience level breakdown for US-published nonfiction print books published in 1923 and 2000.

Table 8. Audience level breakdown, US-published nonfiction print books (1923 and 2000)*

	Schooler	General	Scholar
1923	1 percent	22 percent	77 percent
2000	8 percent	54 percent	38 percent

** For manifestations where it was possible to calculate an audience level. Fraction of manifestations published in 1923 with audience level calculation: 92 percent; in 2000: 92 percent.*

The results in Table 8 suggest a significant shift between 1923 and 2000 in the distribution of print books across juvenile, nonspecialist, and specialist audiences. In 1923, more than three-

quarters of the print books are aimed at a specialist, "scholarly" audience; in contrast, less than 40 percent of the print books published in 2000 fall into the Scholar category, while more than half fall within the purview of a general readership. As with the differences in the distribution across subject categories over time, it is difficult to determine whether this difference in audience level proportions between 1923 and 2000 is a result of shifts in publishing trends, or shifts in collecting and retention decisions by libraries.

The potential influence of collecting and retention decisions raises some interesting speculations on the origins behind the relatively high audience level of older materials in library collections. It may be the case that materials of an academic or scholarly nature have the greatest "survivability" in terms of sustaining their perceived value over time, and therefore exhibit a higher retention rate in library collections than materials aimed at a general or juvenile readership, which may be more likely to be weeded out of the collection over time. If so, a direct correlation would exist between the age of a print book and its audience level. Another explanation relates to the perceived institutional mission of a library in regard to preservation. Academic research libraries tend to have a strong sense of obligation to preserve the cultural and scholarly record. This would suggest that older materials would have a higher likelihood of being retained by academic libraries, compared to other types of libraries. Again, this would tend to support a direct correlation between the age of the book and its audience level; in this case, however, the high audience level may have less to do with the actual intellectual content of the book, and more to do with the preservation decisions of libraries that retain it in their collections.

Focus on Academic Libraries: The "G3"

The analysis in the preceding section examines the aggregate collection of US-published print books in WorldCat, representing the combined print book holdings of libraries of all descriptions: academic, public, special, and so on. Given that many digitization activities, including Google Books, rely heavily on the print book collections of academic libraries, it is useful to take a more focused look at the aggregate collection of print books in WorldCat, emphasizing the holdings of academic libraries. To do this, three large academic research library participants in the Google Books program were selected—one from the East Coast, one from the Midwest, and one from the West Coast. Their print book collections were isolated in WorldCat and then aggregated into one combined collection, with duplicate holdings removed. The result is a reasonably representative illustration of the holdings of academic libraries participating in Google's digitization program, with some rough compensations made for local or regional collecting idiosyncrasies. For the sake of brevity, we will refer to this collection as the Google 3 ("G3") collection.

The G3 collection consists of 9.5 million unique items of all descriptions; of these, about 1.9 million, or 20 percent, are US-published print books. The G3 collection therefore extends over

approximately 12 percent of the overall US-published print book collection in WorldCat. Table 9 presents the distribution of the G3 US-published print book collection over the major time periods impacting copyright status.

Table 9. Distribution of G3 US-published print books in WorldCat, by major US copyright periods

Period	Number	Percentage
Pre-1923	287,246	15 percent
1923-1963	375,637	20 percent
Post-1963	1,224,873	64 percent
Unknown/questionable date	38,128	2 percent

Comparing these results to those in Table 2 suggests that the G3 collection has a slightly higher proportion of materials in the public domain or potentially in the public domain (35 percent) than that exhibited by the overall aggregate collection of US-published print books in WorldCat (31 percent). Indeed, the G3 collection represents a slightly "older" collection than the WorldCat collection, with half of the print books published after 1974 (compared to 1977 for all of WorldCat); two-thirds published after 1961 (compared to 1964 for all of WorldCat); and three-quarters published after 1948 (compared to 1951 for all of WorldCat). The fact that the G3 collection is perceptibly older than the overall WorldCat collection correlates with the speculation in the previous section that the relatively high audience level for older materials may be at least partially attributable to a sustained value of scholarly or specialist materials in certain disciplines over time, and that academic libraries may exhibit a greater willingness to retain older materials in their collections in order to fulfill a perceived obligation to contribute toward the long-term preservation of the scholarly record.

As with the overall WorldCat collection, books that are potentially in copyright constitute the bulk of the G3 collection. More specifically, this consists of all print books in the G3 collection published during or after 1923: 1.6 million print books, or 83 percent of the total.

The G3 collection of in-copyright print books are associated with about 820,000 unique authors (individuals) who served some creative role in regard to one or more of the print book manifestations in the collection, and therefore could potentially hold some form of copyright authority over future use of the book. Table 10 reports the authors associated with the most print book manifestations in the G3 collection of potentially in-copyright books.

Table 10. Authors associated with the most potentially in-copyright G3 print book manifestations

Author	Number of Print Book Titles
William Shakespeare	901
Harold Bloom	739
Bruce Rogers*	644
William Faulkner	554
Marge Piercy	522
Mark Twain	467
Christopher Morley	430
Jacob Neusner	428
John Steinbeck	381
Douglas C. McMurtrie	372

** Note: Bruce Rogers is a noted typographer; due to cataloging conventions, he is often included as sharing in the responsibility for the creation of a work.*

As with the overall WorldCat print book collection, the list is dominated by deceased individuals, most of whom are well-known names in literature and therefore likely to be published and republished frequently over time. In contrast to the WorldCat collection, however, the list in Table 10 is notable for its absence of children's authors, which is likely explained by the fact that the G3 collection is solely comprised of the holdings of three academic libraries.

Approximately 93 percent, or 1.5 million, of the US-published print books in the G3 collection published during or after 1923 are nonfiction, a proportion similar to that found for the WorldCat aggregate collection. Tables 11 and 12 report the subject and audience level breakdown, respectively, for the G3 collection of nonfiction print books.

Table 11. Subject breakdown, G3 nonfiction print books

History and Auxiliary Sciences	12 percent
Language, linguistics and literature	11 percent
Health and medicine	9 percent
Business and economics	9 percent
Engineering and technology	7 percent
Philosophy and religion	5 percent
Art & architecture	5 percent
Sociology	5 percent
Law	4 percent
Education	4 percent
Other	23 percent
Unknown	7 percent

Table 12. Audience level breakdown, G3 nonfiction print books*

Schooler	1 percent
General	21 percent
Scholar	78 percent

** For manifestations where it was possible to calculate an audience level (99 percent of total).*

The breakdown of the G3 nonfiction collection by subject suggests a heavier emphasis by academic libraries on collecting titles in History, Language/Linguistics/Literature, and Health/Medicine than in the corresponding WorldCat collection (see Table 4). More striking, however, is the difference in audience level proportions across the "Schooler," "General," and "Scholar" categories. The results in Table 12 suggest that the G3 collection is strongly oriented toward a scholarly or specialist audience: more than three-quarters of the titles fall into this category, compared to a little over half for corresponding WorldCat collection. This is certainly to be expected, given that the G3 collection represents the holdings of three academic libraries, whose primary function is, of course, to serve researchers and learners. However, given that Google Books and other digitization efforts are heavily engaged with academic libraries and their print book collections, it also suggests that current digitization activities may be on a path to produce a resource predominantly of interest to researchers and students, rather than a general readership.

Conclusion

This article characterizes the aggregate collection of US-published print books in WorldCat, with a special emphasis on materials published during or after 1923, and therefore either

potentially or definitely in copyright. Findings from the analysis indicate that the collection of US-published print books in WorldCat is quite large, encompassing about 15.5 million print books. Nearly two-thirds of these—those published after 1963—have a high likelihood[13] of being in copyright; less than 15 percent—those published prior to 1923—are almost certainly in the public domain, with the rest—those published between 1923 and 1963—potentially in copyright if copyright was renewed. The post-1923 materials collectively account for more than 80 percent, or about 12.6 million, of the US-published print books in WorldCat. It is difficult to predict how many of these print books might be orphan works, but even a small fraction would, in terms of absolute numbers, be considerable, and require a substantial effort to investigate and clear copyright. One study, based on an examination of a random sample of books, estimates a cost of approximately $200 for each title for which digitization and access permissions were obtained.[14]

Analysis of the post-1923 print books in WorldCat suggests significant limitations to automated assessment of copyright status using bibliographic data. Difficulties arise in operationalizing apparently simple concepts: the simple assertion *"this book was first published in the United States"* can be challenged in terms of the definitions of "book" and "published"; uncertainty can even exist over a book's original country of publication.[15] More generally, assertions that we might like to make about information resources in the context of new issues and questions are not always easily generated from existing data sources built for other purposes. While automated analysis of bibliographic data is useful for establishing the general contours of a large collection of print books in terms of copyright status, it is likely insufficient for making a definitive assessment of any one book's copyright status. Manual intervention will almost certainly be required in many cases, especially if the book turns out to be an orphan work.

Investigations aimed at determining copyright status are becoming more prominent in the procedures and workflows of libraries and other organizations. A recent OCLC Research report found that even as these investigations become more common, much ambiguity still surrounds this work in regard to reliable sources of copyright evidence, procedural due diligence, and benchmarks for decision-making.[16] Often, no single source of information exists to establish an item's "copyright provenance," and institutions invoke different rules and criteria for arriving at a copyright status assessment. At this point, copyright investigations seems to be more *ad hoc* than formulaic, more art than science, and oriented toward minimizing risk rather than achieving certainty. The labor intensity—and by extension, the time and expense—associated with copyright investigations underscores the importance of finding ways to reduce costs: for example by sharing the results of copyright investigations to reduce duplicative effort.

Another important finding from the analysis is the prominence of academic institutions as both suppliers and consumers of mass digitization activities like Google Books. From a supply-side perspective, well over half of the total holdings attached to the 15.5 million US-

published print book manifestations in WorldCat belong to academic institutions, indicating that institutions of this kind will necessarily be important sources of the raw materials—print books—needed to supply mass digitization activities. Indeed, most of the current participants in the Google Books library program are academic institutions. From a demand-side perspective, the nature of the materials residing in the collections of academic institutions are, of course, tailored to fit the needs of a research- or specialist-oriented audience, as evidenced by the audience level calculations for the "G3" nonfiction print book collection. Digitization activities operating primarily on the print book holdings of academic institutions will produce digital resources predominantly of interest to academic audiences.

Copyright and regulatory regimes define the limits of what can be done with an information resource. Computing and network technologies afford much greater opportunity to replicate, distribute, access, and repurpose information, and as a consequence, views on what these limits should be have been subject to much wider interpretation. Debate over initiatives like the proposed Google book settlement will help shape these limits, but an important element of the discussion is a thorough understanding of the scope and characteristics of the in-copyright materials in library collections.

Acknowledgments

The authors thank our OCLC Research colleague Jenny Toves for the data on unique authors presented in Tables 3 and 10. Thanks also to Peter Hirtle and Michael Cairns for reading an earlier draft of this article, and providing many helpful comments and suggestions.

Notes and References

1. One of the key components of the Google book settlement is a mechanism for Google to provide for-fee access to digitized copies of in-copyright, out-of-print books. A portion of the fees collected will be allocated to the Book Rights Registry, which would be responsible for distributing them to rights holders. For an overview of the Google book settlement, visit the informational web site http://www.googlebooksettlement.com/. For a summary of the settlement's possible implications for libraries, see: Erway, Ricky. 2009. *Impact of the Google Book Settlement on Libraries* (Revised Version). Dublin, Ohio: OCLC Research. http://www.oclc.org/programs/publications/reports/2009-01.pdf.
2. Lavoie, Brian, Lynn Connaway and Lorcan Dempsey. 2005. "Anatomy of Aggregate Collections: The Example of Google Print for Libraries" *D-Lib Magazine* 11(November/December). doi:10.1045/september2005-lavoie.
3. United States Copyright Office. 2006. *Report on Orphan Works: A Report of the Register of Copyrights.* Washington, DC: Library of Congress. January. p. 15. http://www.copyright.gov/orphan/orphan-report-full.pdf.
4. Works published outside the US are also covered by US copyright law. However, the rules for determining US copyright status for works published overseas are different than those pertaining to works published in the US. To avoid confusion in the analysis which follows, this study is confined to US-published print books only. This is not to diminish the importance of works published outside

the US in digitization activities; indeed, an interesting follow-up study could focus on the scope and characteristics of in-copyright print books published overseas.

5. For consistency, this article uses the same definition of "book" as our earlier article on the "Google 5" (see note 2 above). Books are defined as monographic language materials. Operationally, in the context of a MARC21 record, a book is identified by the codes "a" and "m" in bytes 6 and 7 of the record leader, respectively. Records describing books in print format were identified by eliminating all non-print formats, such as digital, microform, Braille, and so on. Theses/dissertations and government documents are excluded from the analysis, since these materials are usually acquired and managed as separate segments of the library collection. Theses and dissertations could be an important part of digitization efforts like Google Books and may warrant separate study. US copyright law often treats government documents differently than other publications; for example, a work produced by the federal government is generally considered in the public domain upon publication.
6. The term manifestation is formally defined in the FRBR model. See: IFLA Study Group on the Functional Requirements for Bibliographic Records. 2009. *Functional Requirements for Bibliographic Records.* p. 21. http://www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf.
7. See http://www.diglib.org/forums/spring2009/presentations/HathiTrust.pdf. HathiTrust also found that of the 2.8 million volumes digitized as of April 2009, about 15 percent were in the public domain.
8. See Ringer, Barbara. 1961. *Study No. 31: Renewal of Copyright*. Washington, DC: US Government Printing Office. (reprint).
9. Author names were extracted from the MARC 100 and 700 fields. Figures reported include personal names only; corporate authors are not included.
10. Subject categories are adapted from the OCLC Conspectus divisions. For more information, see: http://www.oclc.org/support/documentation/collectionanalysis/using/introduction/introduction.htm#conspectus_WCA.
11. To learn more about the audience level metric, see: http://www.oclc.org/research/activities/audience/default.htm. See also: O'Neill, Edward, Lynn Connaway and Timothy Dickey. 2008. "Estimating the Audience Level for Library Resources." *Journal of the American Society for Information Science and Technology.* 59,13: 2042-2050. Pre-print available online at: http://www.oclc.org/research/publications/archive/2008/oneill-jasist.pdf.
12. Note that the 2007 cut-off date refers to the time the book's record was entered into WorldCat, not when it was published. It is possible, for example, that a book published in 1995 would not get entered into WorldCat until several years later.
13. The qualifier "high likelihood" is included because materials published before 1989 required a copyright notice to be considered "in copyright." After 1989, the copyright notice requirement was dropped. Therefore, we must assume that some fraction of post-1963 US-published books were published without a copyright notice, and would therefore be in the public domain.
14. Covey, Denise Troll. 2005. *Acquiring Copyright Permission to Digitize and Provide Open Access to Books* (Washington, DC: CLIR). http://www.clir.org/pubs/reports/pub134/pub134col.pdf.
15. Hirtle, Peter B. (2008) "Copyright Renewal, Copyright Restoration, and the Difficulty of Determining Copyright Status" *D-Lib Magazine.* 4 (7/8). doi:10.1045/july2008-hirtle.
16. Proffitt, Merrilee, Arnold Arcolio and Constance Malpas. 2008. *Copyright Investigation Summary Report*. Dublin, Ohio: OCLC Research. http://www.oclc.org/programs/publications/2008-01.pdf.

Cloud-sourcing Research Collections:
Managing Print in the Mass-digitized Library Environment

Constance Malpas

Into being
The clouds condense, when in this upper space
Of the high heaven have gathered suddenly,
As round they flew, unnumbered particles—
World's rougher ones, which can, though interlinked
With scanty couplings, yet be fastened firm,
The one on other caught.

Lucretius *De rerum natura*, Book V
trans. William Ellery Leonard (1921)

Acknowledgments

The Cloud Library project emerged out of a series of discussions that began with Carol Mandel, Jim Neal, John Wilkin and Jim Michalko in 2009. These individuals provided leadership and vision that guided all the work that followed.

Library staff from New York University, Columbia University, the New York Public Library and Princeton University participated in a variety of meetings, conference calls and e-mail exchanges that helped to give shape to the project. The Andrew W. Mellon Foundation contributed financial support under a grant ably administered by Chuck Henry at the Council on Library and Information Resources (CLIR).

Michael Stoller, Bob Wolven, Zack Lane, Matthew Sheehy, Marvin Bielawski and Eileen Henthorne made essential contributions to the project, not least in helping to compile ReCAP holdings data for inclusion in our analysis. Kat Hagedorn and Jeremy York provided expert technical and operational support from Hathi. Jenny Toves ensured that WorldCat data extractions were available on schedule.

This paper was originally published in January 2011 as an OCLC Research report.
http://www.oclc.org/research/publications/library/2011/2011-01.pdf.

I am grateful to Jim Michalko, John Wilkin and Paul Courant for their many thoughtful questions and suggestions about the data analysis and interpretation. Lorcan Dempsey and Brian Lavoie also provided insights and helpful methodological guidance along the way.

Particular thanks are due to Roy Tennant and Bruce Washburn, who provided expert programming support over the course of this project and routinely produced small miracles, and to Patrick Confer for his diligent editorial work in preparing the final report.

Executive Summary

The Cloud Library project was jointly designed and executed by OCLC Research, the HathiTrust, New York University's Elmer Holmes Bobst Library, and the Research Collections Access & Preservation (ReCAP) consortium, with support from The Andrew W. Mellon Foundation. The objective of the project was to examine the feasibility of outsourcing management of low-use print books held in academic libraries to shared service providers, including large-scale print and digital repositories.

The following overarching hypothesis provided a framework for our investigation:

- The emergence of a mass-digitized book corpus has the potential to transform the academic library enterprise, enabling an optimization of legacy print collections that will substantially increase the efficiency of library operations and facilitate a redirection of library resources in support of a renovated library service portfolio.

From this, a number of research questions emerged:

- What is the scope of the mass-digitized book corpus in the HathiTrust Digital Libray and to what degree does it replicate print collections held in academic research libraries?
- Can public domain content in the HathiTrust Digital Library provide a suitable surrogate for low-use print collections in academic libraries?
- Is there sufficient duplication between shared print storage repositories and the HathiTrust Digital Library to permit a significant number of academic libraries to optimize and reduce total spending on local print management operations?
- What operational gains might be obtained through a selective externalization of collection management activities?

Based on a year-long study of data from the HathiTrust, ReCAP, and WorldCat, we concluded that our central hypothesis was successfully confirmed: there is sufficient material in the

mass-digitized library collection managed by the HathiTrust to duplicate a sizeable (and growing) portion of virtually any academic library in the United States, and there is adequate duplication between the shared digital repository and large-scale print storage facilities to enable a great number of academic libraries to reconsider their local print management operations. Significantly, we also found that the combination of a relatively small number of potential shared print providers, including the Library of Congress, was sufficient to achieve more than 70% coverage of the digitized book collection, suggesting that shared service may not require a very large network of providers.

Analysis of the distribution of subject matter and library holdings represented in the HathiTrust Digital Library and shared print repositories further confirmed that the digital corpus is largely representative of the collective academic library collection, suggesting a broad potential market for service. A further positive finding was that monographic titles in the humanities constitute the greatest part of the mass-digitized resource, which may indicate that some relatively under-resourced disciplines will begin to benefit from a digital transformation that has already powered enormous innovation in the sciences. As detailed below, we also found that substantial library space savings and cost avoidance could be achieved if academic institutions outsourced management of redundant low-use inventory to shared service providers.

Our findings also revealed some important obstacles and limitations to implementing changed print management practices in the current library operating environment. The following are among the most important constraints we identified:

- The proportion of public domain content in the HathiTrust Digital Library is relatively small (approximately 16% of titles in June 2010) and typically represents material that is not widely held in the library system; as a result, the number of libraries that might hope to reduce local print management costs for these titles through negotiated agreements with the HathiTrust and shared print providers is quite low. Moreover, the age and subject distribution of titles in the public domain is not representative of academic research collections as a whole. In sum, the public domain corpus as currently defined by U.S. copyright law cannot be considered a viable surrogate for any academic print collection.

- While significant duplication was found between the HathiTrust Digital Library and multiple large-scale library storage collections, it was apparent that no single print storage repository could offer coverage sufficient to enable significant space savings or cost avoidance for a given client library. Put another way, effective shared print storage solutions will depend upon a network of providers who will need to optimize holdings as a collective resource.

- The absence of a robust discovery and delivery service based on collective print storage holdings is an impediment to changed print management strategies, especially for digitized titles in copyright.

It is our strong conviction, based on the above findings, that academic libraries in the United States (and elsewhere) should mobilize the resources and leadership necessary to implement a bridge strategy that will maximize the return on years of investment in library print collections while acknowledging the rapid shift toward online provisioning and consumption of information. Even, and perhaps especially, in advance of any legal outcome on the Google Book Search settlement, academic libraries have a unique opportunity to reconfigure print supply chains to ensure continued library relevance in the print supply chain. In the absence of a licensing option, online access to most of the digitized retrospective literature will be severely constrained. Demand for print versions of digitized books will continue to exist and libraries will be motivated to meet it, but they will need to do so in more cost-effective ways. In the absence of fully available online editions, full-text indexing of digitized in-copyright material provides a means of moderating and tuning demand for print versions and should facilitate the transfer of an increasing part of the print inventory to high-density warehouses. Viewed in this light, shared print storage repositories could enable a significant and positive shift in library resources toward a more distinctive and institutionally relevant service portfolio.

Our study assessed the opportunity for library space saving and cost avoidance through the systematic and intentional outsourcing of local management operations for digitized books to shared service providers and progressive downsizing of local print collections in favor of negotiated access to the digitized corpus and regionally consolidated print inventory. As detailed in the report that follows, the organizational change required to achieve these gains is likely to be substantial and challenging to implement. Yet, the opportunity costs of inaction may prove even greater than the risks of enacting shared print management regimes. Many of the positive transformations that academic library directors hope to achieve in the next decade or so will require a fundamental shift in collections management. The scope and scale of change that is possible may be judged by these key findings:

- As of June 2010, the median rate of duplication between titles held by university libraries in the U.S. Association of Research Libraries (ARL) and the HathiTrust Digital Library exceeds 30%; that is to say, nearly a third of the content purchased by research-intensive libraries in the United States has already been digitized and is preserved in a shared digital repository.

- If the current growth trajectory of the HathiTrust Digital Library is sustained, we can project that more than 60% of the retrospective print collections held in ARL

libraries will be duplicated in the shared digital repository by June 2014. This growth rate far exceeds average annual acquisitions in ARL libraries, suggesting that the digital replication of legacy collections will outpace growth of new physical collections, enabling a transformation in traditional library operations, staffing and space requirements.

- The median space savings that could be achieved at an ARL library if a robust shared print offer were in place today amounts to approximately 36,000 linear feet or the equivalent of more than 45,000 assignable square feet (ASF). These are conservative estimates based on the assumption that holding libraries own a single copy of each duplicated title. Actual space savings could be much greater. In practical terms, this means each library could recover space sufficient for a learning or research commons, media lab, or office space for faculty and visiting scholars.

- The total annual cost avoidance that could be achieved if shared print service provision for mass-digitized books were available today would amount to a figure between $500,000 and $2 million per ARL library, depending on the physical environment (e.g., open stacks on campus or high-density off-site storage) in which the titles would be managed locally.

Academic library directors can have a positive and profound impact on the future of academic print collections by adopting and implementing a deliberate strategy to build and sustain regional print service centers that can meet aggregate demand with aggregate supply. Beyond the obvious operational efficiencies of consolidating low-use, digitized print volumes into shared service collections there is an important strategic advantage to reconfiguring collective inventory that is increasingly devalued as an institutional asset. A proactive effort to rationalize collections that are undergoing a radical phase change from print to digital will enable libraries to achieve a careful and measured wind-down of operations that no longer deliver distinctive value, while continuing to uphold a vital preservation and access mandate.

The shared infrastructure needed to support a broad-based externalization of legacy print management functions is unlikely to emerge without directed action and decision-making by leaders in the academic library community. Individuals and organizations interested in advancing these changes are encouraged to consider the following recommendations:

Library directors and managers can . . .

- Advocate in favor of licensed access to the mass-digitized resource as part of a comprehensive strategic plan in which the library can reassert its role as a vital part of the academic enterprise.
- Engage directly with faculty and academic officers to communicate a compelling strategy in which selective externalization of traditional functions is demonstrably improving the institution's ability to fulfill an academic and research mission.
- Support the HathiTrust's ongoing efforts to expand public access to the mass-digitized book corpus by affiliating with the organization as a content contributor or sustaining partner.

Prospective shared print providers, including managers of large print storage facilities, can . . .

- Proactively build collections that will deliver maximum operational value to external audiences; leverage the collective library investment in mass digitization and the HathiTrust by accelerating the transfer of mass-digitized titles to print preservation repositories.
- Contribute to the establishment of a common service profile by surfacing model agreements and engaging in community dialog about the operational and business requirements of shared service provision.

Research organizations, including OCLC Research, Ithaka S+R, JISC and other similar entities, can . . .

- Advance our collective understanding of the changing profile of demand for legacy print collections in the mass-digitized environment.
- Help to characterize the optimal redistribution of library resources in different regional and national contexts.

Funding bodies, including IMLS, the Mellon Foundation, NEH and others, can . . .

- Provide funding to support the implementation of shared print management through grants to libraries and other organizations to subsidize the direct costs of title selection and processing until such activities are fully subsumed as ongoing library operations.

Introduction

In spring 2009, a group of ARL directors came together to discuss a common set of challenges and opportunities facing university libraries and identify some shared strategies for responding to them. A number of circumstances were converging that appeared to offer some potential relief from critical space pressures in the library and the increasingly burdensome operations associated with managing a large local inventory of low-use print collections.

The seemingly imminent resolution of the Google Book Search settlement was an important motivating factor: academic libraries were confronting the prospect, at once daunting and liberating, of licensed access to a massive aggregation of digitized books from major U.S. research collections. Would such a collection substantially duplicate local print holdings? If so, what consequences might ensue for traditional academic library operations?

At the same time, the emergence of the HathiTrust, a shared digital repository consolidating much of the library-contributed content from the Google Books database, appeared to resolve many of the concerns the library community had regarding long-term stewardship of the mass-digitized book corpus. In combination with the large aggregations of low-use print collections managed in high-density library storage facilities, Hathi might bridge the gap between a well-documented decline in the use of academic print collections and the anticipated shift toward scholarly reliance on full-text electronic resources.

The fact that critical elements of the shared infrastructure needed to effect a large-scale transition from print to electronic research collections were owned and managed by the library community itself gave library directors confidence that the timing and outcomes of this transition could be managed according to the needs of the academic community and not dictated by the business objectives of commercial providers. Were the combined resources of Hathi and large-scale shared print providers already sufficient to mobilize a change in library operations? What was the scope of service likely to be? How much and what kind of value would it need to deliver? Who—which kinds of libraries and in what number—would benefit? These questions were compelling enough to justify a joint research project in which potential service providers and consumers could explore business requirements, service expectations and feasibility of implementation.

The initiative that emerged from these discussions within ARL came to be known as the "Cloud Library" project, because it posited a future in which library collections and services would be sourced from external providers, reducing local infrastructure and operational expenditures in a manner analogous to the cloud-sourced business and computing solutions that now prevail in the commercial and high-tech sectors. Funded by The Andrew W. Mellon Foundation, the project was staffed by a team of investigators from the HathiTrust, the

Research Collections Access and Preservation consortium (ReCAP), New York University Libraries, and OCLC Research. This report provides a high-level summary of findings from this project.

Premise

The research questions that motivated this study reflect a conviction shared by all of the participating institutions: ***the emergence of a mass-digitized book corpus has the potential to transform the academic library enterprise, enabling an optimization of legacy print collections that will substantially increase the efficiency of library operations and facilitate a redirection of library resources in support of a renovated library service portfolio***. We started from the presumption that academic libraries will be motivated to transfer resources (space, personnel, and capital) from local print management operations to shared print and digital repositories in proportion to the tangible benefit that cooperative management confers. We were therefore less interested in examining the theoretical advantages of shared service provision than in characterizing the operational gains (space recovery and cost avoidance) that might be obtained through a selective externalization of collection management activities.

Methodology

Between June 2009 and June 2010, a monthly snapshot of records was harvested by OCLC Research from the publicly available HathiTrust metadata repository. These records were machine-processed to extract OCLC numbers and, where necessary, to extract and map alternative identifiers (LCCN, ISBN or ISSN) to valid OCLC numbers. The resulting batch of OCLC numbers was used to extract bibliographic records and holdings data from the WorldCat database each month. These bibliographic master records were then merged with selected Hathi metadata and (starting in September 2009) a sample of associated ReCAP repository customer codes to produce a single, consolidated dataset for analysis.

A master database was built to support analysis of the compiled data, which was programmatically enhanced to support analysis of key attributes of the aggregate collection, including broad subject areas, total library holdings, institutional source of the digitized text and copyright status. This database was enriched each month with successive snapshots of the Hathi repository, mapped to WorldCat holdings and ReCAP customer codes as described above. By June 2010, the project database comprised 37 million records, representing a longitudinal view of the growing corpus of library-owned titles that are duplicated in print and digital repositories.

Scope of Analysis

In the twelve months covered by this project, the HathiTrust Digital Library doubled in size, increasing from approximately 3 million volumes to more than 6 million volumes. ***On a per-volume basis, the shared digital repository is now larger than the average ARL library collection***; the median reported holdings at university-based ARL libraries in 2008 was approximately 3.5 million volumes. Because our analysis of the HathiTrust collection focuses on unique titles (manifestations or editions), rather than physical items, the number of records we compiled each month was somewhat smaller than the number of records in the Hathi metadata repository. Not every volume in the HathiTrust represents an individual book or journal title, and there is at least some duplication in content ingested from different contributors; as a result, the total number of volumes in the Hathi repository is more than the number of titles covered in our analysis. In June 2009, we identified approximately 2 million unique titles in the HathiTrust Digital Library; by June 2010, that number had grown to more than 3.6 million titles. For purposes of comparison, this represents ***a collection comparable in scope to research libraries in the top tier of the U.S. ARL rankings***, based on holdings set in the WorldCat database. Indeed, at the time of writing, the number of unique titles in the HathiTrust Digital Library exceeds the number of titles cataloged and held by many research libraries.

A key goal of this research project was to assess the scope of coverage in shared print and shared digital repositories, with a view to understanding how the combined resources might enable a local reduction in redundant print inventory. For this reason, it was important to understand how much of the print storage collection in ReCAP is duplicated—or is likely to be duplicated—in the HathiTrust Digital Library. As of this writing, the shared ReCAP facility holds more than 8 million items contributed by the three partner libraries. Since the ReCAP collection is not currently visible as a discrete set of holdings in WorldCat, and building a union catalog of ReCAP holdings was beyond the scope of this project, we based our analysis on a representative sample of ReCAP holdings supplied by Columbia University and NYPL. Taken collectively, Columbia and NYPL's ReCAP holdings amount to more than 75% of current inventory and this was deemed to be sufficient for our analysis.

The sample supplied to us included a broad range of materials managed under 14 different ReCAP customer codes, each representing a different set of request and circulation rules. The large size and broad scope of the sample gave us reasonable confidence that findings from our analysis could be generalized across the ReCAP collection as a whole. Storage, selection and transfer protocols at the three partner libraries are based on common parameters (low use monographs; journals duplicated in electronic format), so that the nature, if not the content, of the materials contributed by each is likely to be comparable.

To provide a baseline against which duplication of ReCAP holdings in the HathiTrust Digital Library might be assessed, we periodically compared patterns in the ReCAP sample against other large-scale print storage collections that are more readily subject to analysis in WorldCat. Findings from these analyses are presented below.

Summary of Findings

In this section, the scope and character of holdings in the HathiTrust Digital Library and ReCAP print repository are examined with a view to their potential value in a shared service environment. We first consider the range of holdings in the HathiTrust Digital Library, on the premise that the vast and still expanding scope of the mass-digitized corpus will be a key driver in the transformation of academic library collections and services. We then examine the intersection of titles held in the HathiTrust Digital Library and the ReCAP print repository to assess the degree to which large-scale storage collections might serve as print management hubs, reducing the total cost of preservation and access for low use print resources. Finally, we explore how this shared infrastructure might affect library operations and resource allocations in a research-intensive academic library, using NYU's Elmer Holmes Bobst library as an exemplar.

Shared Digital Repository Profile: HathiTrust

Over the period of study, the number of volumes in the HathiTrust Digital Library more than doubled, growing from about 3 million items to more than 6.3 million items; the number of titles increased by 90%, from just over 1.9 million titles in June 2009 to about 3.64 million titles in June 2010. Growth was variable from month to month, ranging from a low of about 43,000 new titles in April 2010 to a high of more than 297,000 new titles in November 2009. On average, the number of unique titles in the database increased by about 6% each month. This represents an average increase of nearly 150,000 new titles each month. The ratio of volumes to titles in the repository remained relatively stable at 1.6:1 over the twelve months of this study.

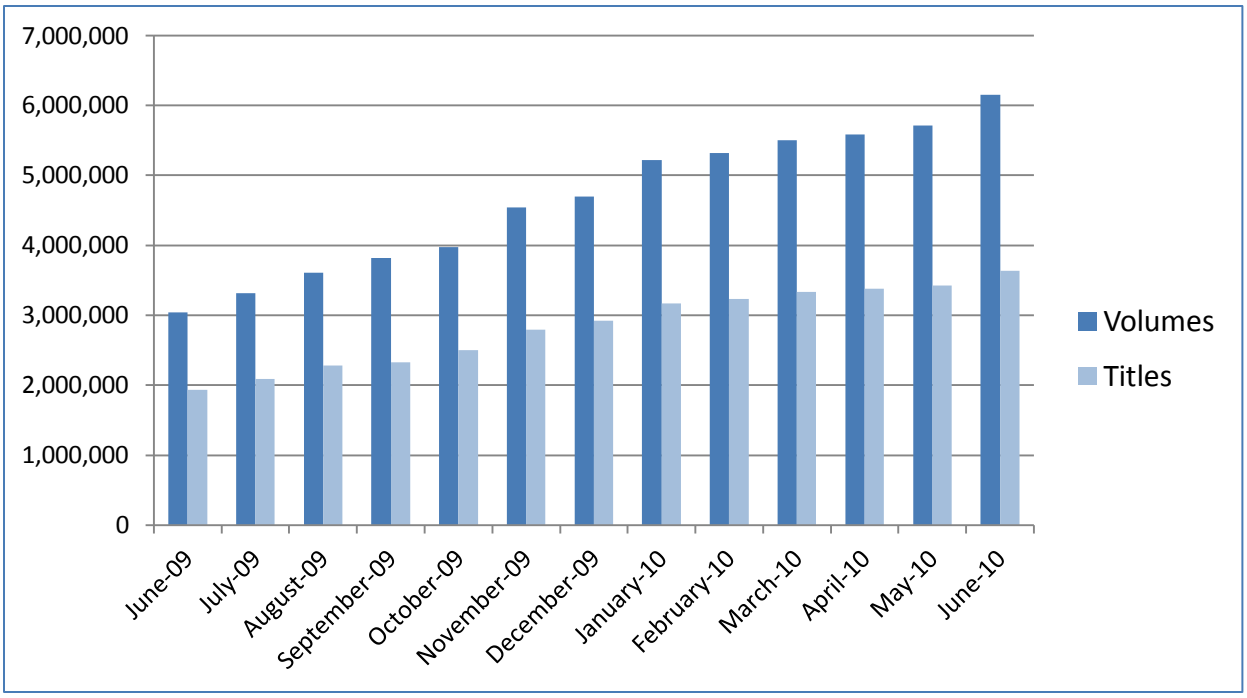


Figure 1. Growth of HathiTrust Digital Library collection (June 2009-June 2010)

If this rate of growth is sustained, we can expect the HathiTrust Digital Library to rival major research library collections in both size (volumes) and scope (titles) in a matter of a few years. Based on the projections shown below, we can anticipate that the ***HathiTrust Digital Library collection may be equal in size to Harvard University Libraries*** (which reported holdings of some 16 million volumes in the 2007-2008 ARL Annual Statistics) ***by 2013. Within a decade, it could cross the threshold of 30 million volumes, making it larger than the U.S. Library of Congress is today***.

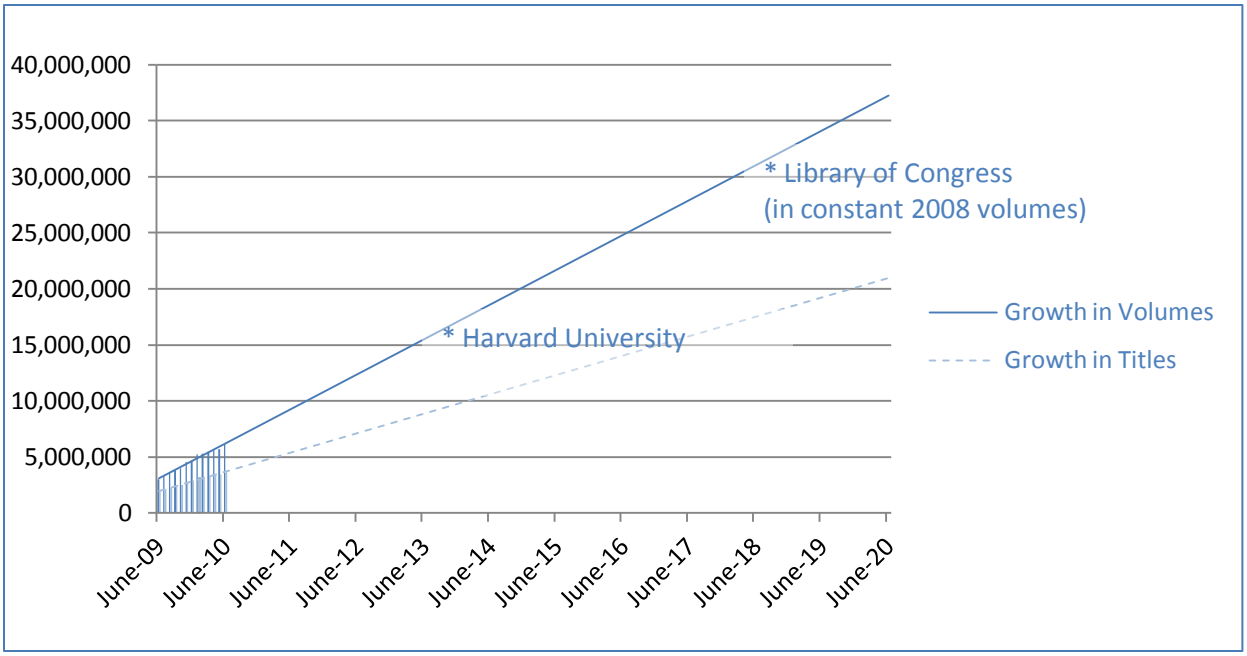


Figure 2. Projected growth of HathiTrust Digital Library (June 2010–June 2020)

For ease of presentation, these projections compare the growth of Hathi to a baseline of constant volume counts at the largest university and non-university ARL collections. Of course, it is reasonable to expect that volume counts for print holdings at these libraries will continue to grow over the next decade; however, the current growth rate of the HathiTrust Digital Library substantially outpaces median annual growth rates at ARL member libraries (approximately 2% of total volume count, based on recent ARL statistics) so we can anticipate that the overlap in digitization of retrospective print holdings will continue to grow faster than the acquisition of new print titles.[1]

Understanding the relative distribution of document types in the HathiTrust Digital Library archive is important to characterizing and quantifying its value as a potential surrogate to locally held academic library print collections. Since the advent of the e-journal transition of the 1990s, university libraries have regarded print versions of dual-format titles as obvious targets for relegation to storage facilities. A major focus of the present study was to determine the degree to which mass digitization of library print collections has resulted in the creation of a digitized book corpus sufficient to enable a similar shift in management of monographic holdings. It is not yet known if the emergence of a large-scale digital book corpus will be sufficient to effect a change in scholarly practice comparable to what has been achieved in the transition from print to electronic journals. Nor is it possible to foresee when, or even if, a legal settlement will be reached that will permit Google to offer universities licensed access to the millions of books that have already been digitized through its partnerships with academic libraries. While uncertainty about the speed and timing of the

format transition for scholarly monographs abounds, we can at least begin to assess the scope and coverage of the academic print collection as it is mirrored in the mass-digitized corpus preserved in the HathiTrust Digital Library.

Document Types

A vast majority of titles in the Hathi repository represent monographic language-based materials (books). Based on our analysis, ***books account for 95% of all titles in the HathiTrust Digital Library*** for which we were able to identify an OCLC number; serial titles comprise approximately 4% of such titles. The remainder of the archive is composed of digitized musical scores, articles, visual resources and the like. While the total volume of nonbook and nonjournal titles in the archive, as measured in absolute numbers, is impressive (amounting to nearly 50,000 titles in June 2010), these materials collectively represent only about 1% of the Hathi corpus.

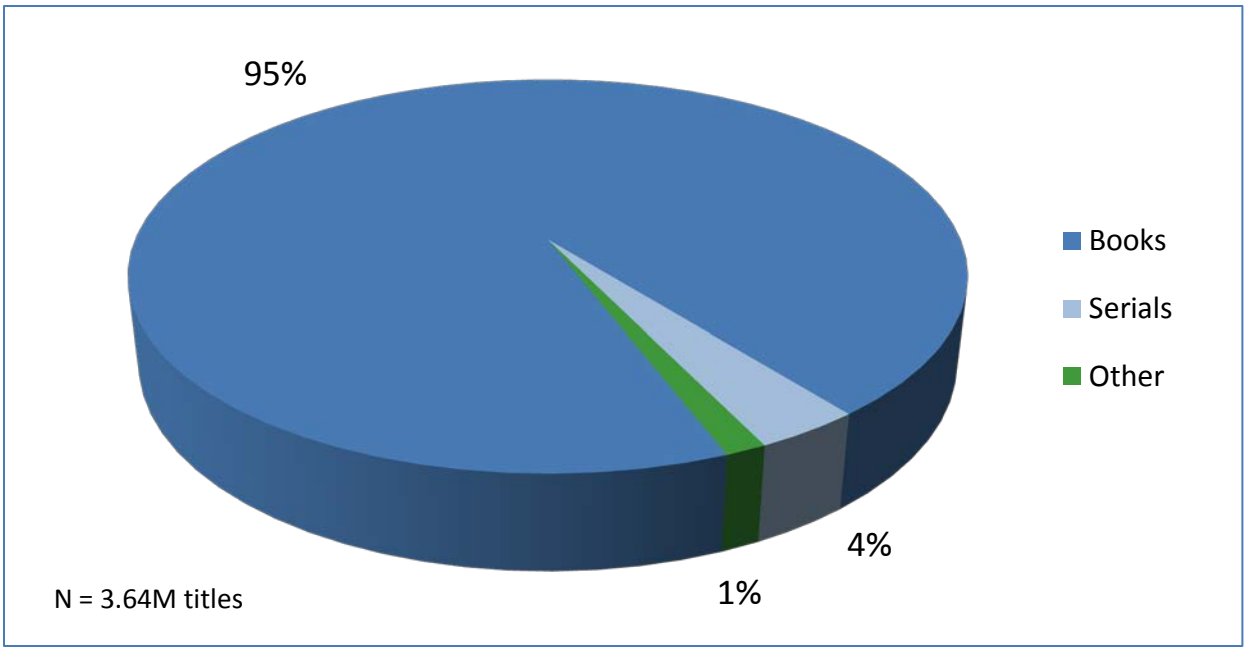


Figure 3. Primary document types of titles in HathiTrust Digital Library (June 2010)

Over the course of our study, an increase in the diversity of document types in the HathiTrust Digital Library has been noted, as indicated by a slight but perceptible shift in proportional distribution of titles. Between June 2009 and June 2010, the relative volume of "other" document types increased from a tenth of a percent (.1%) to a third of a percent (.3%) of all titles in the database. As of June 2010, musical scores account for the vast majority of titles in this "other" category. It is not certain what the impact of this trend is likely to be, but one

might speculate that a sustained growth in nonbook and nonserial titles will be associated with a net decrease in the number of libraries eligible to transfer preservation functions to Hathi, as aggregate library holdings for nonbook materials tend to be significantly lower than for book and "book-like" materials. Based on an August 2010 snapshot of the WorldCat database, for example, the average number of library holdings set on an individual monographic title is nine; for musical scores, by contrast, the average number of holdings is four. A shift towards greater representation of nonbook and journal content in the archive may meet the needs of current contributors, but it is not likely to support a broader externalization of preservation functions in other libraries.

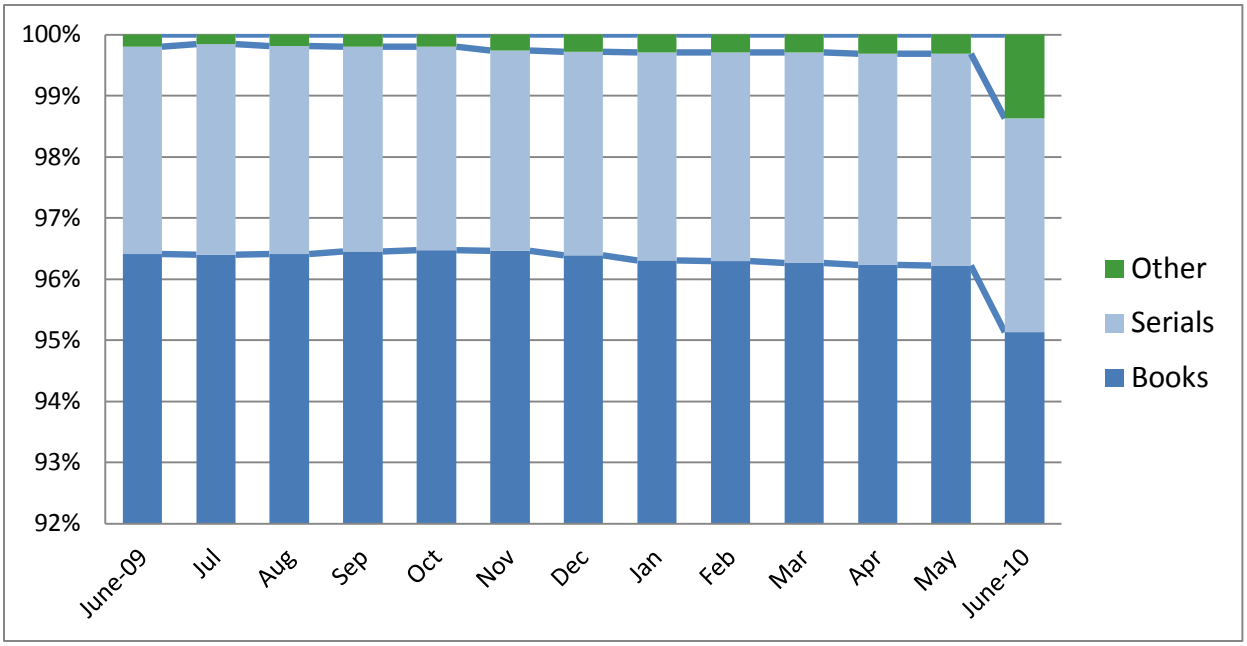


Figure 4. Distribution of HathiTrust Digital Library titles by document type (June 2009–June 2010)

Because we are primarily concerned with assessing the potential impact of shared digital and print archives on library-managed print collections, and because books continue to represent the single largest cost driver in library operations, the analysis that follows focuses on books and not other library-owned material types.

Subject Distribution

Individual titles in our dataset were coded with broad and narrow topical descriptors derived from the OCLC Conspectus subject classification.[2] We analyzed the frequency of these codes to determine which subject areas predominate in the digitized Hathi corpus, with the expectation that libraries will adjust print retention policies in view of differing disciplinary reliance on physical books. As shown in the chart below, more than 50% of titles in the

HathiTrust Digital Library in June 2010 represent content from traditional humanities fields: language and literature, history, philosophy, art and architecture, etc.

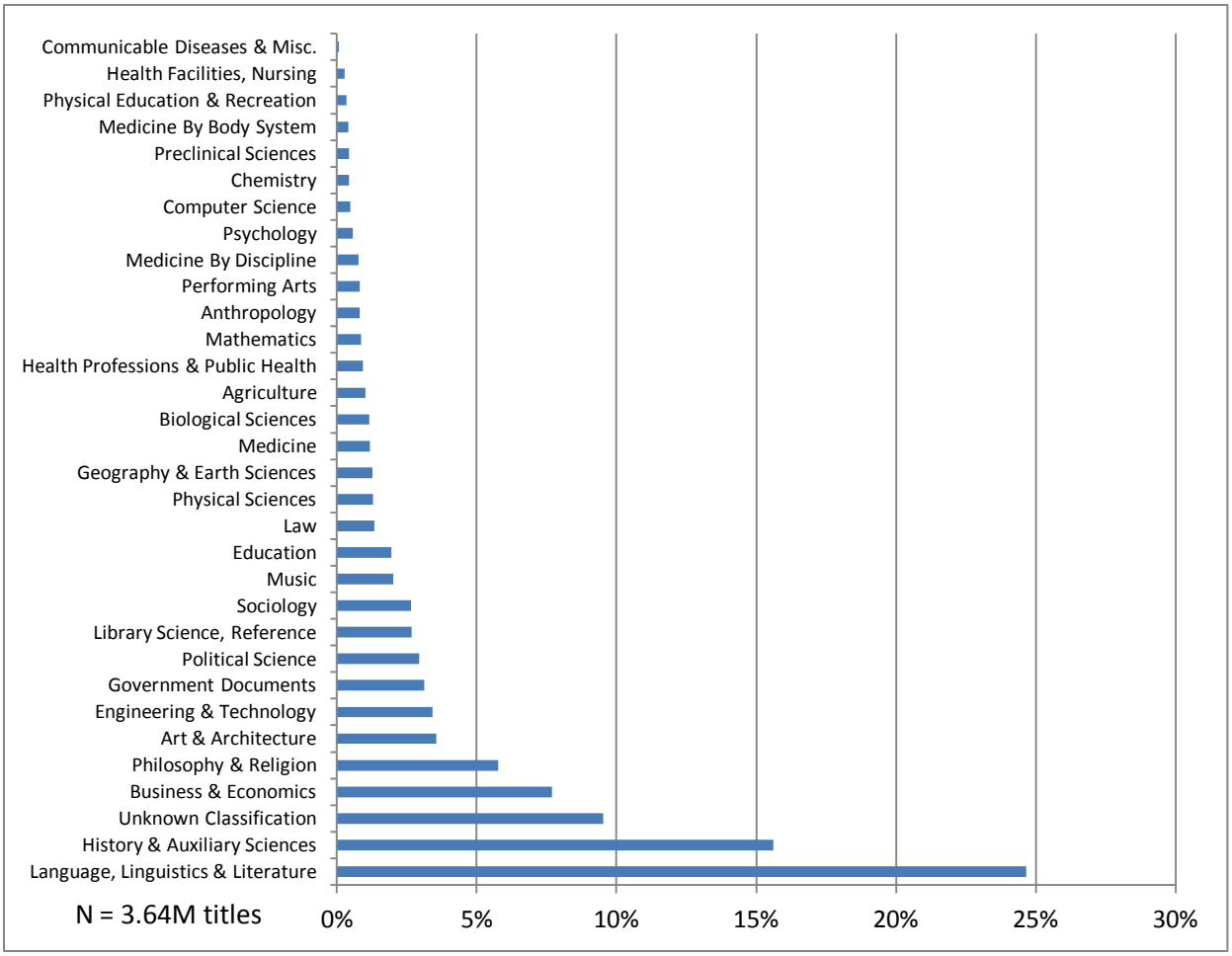


Figure 5. Subject distribution of titles in HathiTrust Digital Library (June 2010)

The ***relative abundance of titles in the humanities*** (history, language and literature, philosophy) ***in the HathiTrust Digital Library provides encouraging evidence that mass digitization of library book collections is redressing a long-observed imbalance in the online availability of scholarly resources in the humanities and social sciences***, compared to the natural sciences and technology. The HathiTrust's explicit mandate to increase the educational and research value of mass-digitized books and to improve public access to them should raise library confidence that the vast and still growing aggregation of digitized texts will not only prove satisfactory to students and researchers, but also sufficiently robust to enable a gradual transformation of the library enterprise, as operations shift from locally managed print to collectively managed digital formats.

Books in the humanities typically constitute a significant share of any academic library's print inventory. While circulation rates for these materials are generally low, they are commonly considered essential to the practice of research and teaching. They have an equally important symbolic value as the embodiment of institutional investment in disciplinary communities that are comparatively "under-resourced" in higher education. Historians are often among the most vociferous critics of any effort to shift physical collections from a central library location to a peripheral shelving or storage annex. Their unease and sometimes outright hostility to well-intentioned strategies for optimizing the distribution of library collections are motivated by deep and praiseworthy concerns about long-term preservation and access to the scholarly record. Until recently, academic libraries have had few options but to retain as much of this low-use but highly valued material on campus as possible; providing direct and unmediated access to print volumes has been the easiest and sometimes the only way to satisfy faculty expectations. The large-scale format transition achieved through mass digitization of these legacy collections has the capacity to transform academic library operations by expanding the range of access options that are available to faculty and students, while simultaneously enabling library managers to make more strategic use of diminishing collections space.

Though smaller in size, other subject-based categories of content represented in the HathiTrust Digital Library are also worthy of note. For example, library owned reference collections (fact books, annual bibliographies, statistical yearbooks, etc.) amount to more than 95,000 titles in the HathiTrust Digital Library. While this constitutes only 3% of the Hathi collection as a whole, it represents a significant potential cost savings for libraries since superseded reference titles are generally regarded as a low print preservation priority; thus, we can imagine that expectations for redundancy in library holdings for these resources might be significantly impacted by replication in the HathiTrust Digital Library. There are more than 20,000 digitized reference titles in the HathiTrust Digital Library that are held in print format in 100 or more libraries. If redundancy in system-wide holdings were reduced to just 15 print copies per title—a figure that recent studies suggest is adequate to ensure survivability of at least one copy for the next one hundred years (Schonfeld, 2009)—a total of more than 20 miles in shelf space might be recovered by libraries.

Government publications are another category of material for which substantial reductions in library print inventory might be achieved, in view of the preservation guarantees provided by the HathiTrust Digital Library. As of June 2010, there are ***more than 100,000 government documents in the HathiTrust Digital Library collection***. More than 40% of these titles are held by in excess of 100 libraries—far more than is required to support the requirements of the U.S. Federal Depository Library Program, for example, and arguably more than is needed to ensure universal access. Because government publications are typically exempt from copyright restrictions, there is every reason to believe that digitized

versions will be widely available, further reducing the need for print inventory. Among titles classified as government documents in the HathiTrust Digital Library, nearly 80% are designated as public domain content. ***One can easily imagine that many academic libraries will choose to downsize local document collections in favor of online versions; for such institutions, the Hathi preservation services could provide a compelling and cost-effective alternative to local print archiving***. Even those libraries that choose to maintain their status as selective depositories could achieve significant cost savings by transferring physical copies of the government publications replicated in the HathiTrust Digital Library to high-density storage facilities.

Additional research is needed to discover what subject areas are included in the "unknown classification" category; given the large number of titles in question (more than 300,000 as of June 2010), this appears to be a fruitful area for study, especially because—as is noted below—more than 20% of titles in this category are in the public domain. Such analysis was beyond the scope of the present study.

Although it was not a focus of our analysis, we did note the presence of many large FRBR work sets in the HathiTrust Digital Library, which suggests some intriguing possibilities not only for discovery services but also for cooperative management and preservation. Thus a library holding a print version of a low-use, in-copyright title might be more likely move it to a cost-efficient high-density facility if it had negotiated with Hathi to provide a link to a public domain digitized surrogate. Another library might opt to withdraw holdings based on levels of duplication in the HathiTrust Digital Library for the associated work set. Our investigation suggests that 5% or more of titles in the Hathi collection (as of June 2010) can be associated with larger work sets. Popular titles like Defoe's *Robinson Crusoe* or Swift's *Gulliver's Travels*, as well as classics like Lucretius' *De rerum natura* or Homer's *Iliad*, are each represented by hundreds of digitized editions in the HathiTrust Digital Library; the long-term preservation of the intellectual work embodied in these manifestations is, to coin a phrase, virtually guaranteed.

It is worth considering that ***as the number and scope of variant editions in Hathi grows, its value to the academic library community may increase exponentially***, enabling the Trust to offer valuable preservation services even to libraries that have contributed no content to the collection. This could significantly increase the market for Hathi preservation and access services and would entail measuring duplication in holdings not on a volume or title level, but on a FRBR work level. In this scenario, Hathi would provide a bridge to facilitate the transition of scholarly practice from print to electronic resources, incrementally reducing demand for, and expectations of, physical proximity to print holdings. Thus, some number of the more than two thousand libraries that hold print editions of Sinclair Lewis' *Babbitt* might reasonably opt to shift the locally held print version to a high-density storage warehouse

while providing patrons with full-text reading access to a digitized public domain version. Libraries availing themselves of this service would still be "on the hook" for preservation of editions not replicated in the Hathi collection, but could manage those resources more efficiently. In this sense, every library that holds an edition of a work represented in the Hathi repository is in a position to derive some tangible benefit from participation in the network. This has important implications for the future growth of the HathiTrust Digital Library, since the capacity to benefit from participation will increase as the scope of the collection increases to include more widely held titles and work sets.

Rights Status

One of the hypotheses that this study set out to test is that the HathiTrust Digital Library represents a potentially rich source of digital surrogates that might, over time, effectively replace a substantial proportion of low-use print collections in academic libraries. It was therefore important not only to examine the size and growth of this corpus over time, but also to consider the degree to which it replicates print holdings in the wider academic library system.

For most of the twelve-month period covered by this study, the relative proportion of in-copyright and public domain content in the HathiTrust Digital Library remained stable, with about 17% of volumes designated as public domain material. This figure increased to about 20% near the end of the project, due in part to a programmatic change in the HathiTrust rights determination algorithm that affected a large number of items ingested earlier in the year. On a per-title basis, a similar distribution was noted over the course of the study, with about 12% of titles designated as public domain content, rising to approximately 16% by the project's close. As of June 2010, approximately 590,000 titles were designated as "full view" content available for onscreen reading in the HathiTrust platform. About 96% of these public domain titles are books, similar to the distribution pattern noted above for the HathiTrust Digital Library as a whole.

In other respects, the public domain corpus presents significant differences. First and most obviously, titles in the public domain are typically older publications, either published before the 1923 threshold (for U.S. publications) or in the period between 1923 and 1976, when some previously in-copyright titles may be "reborn" as public domain content, either by direct negotiation with the rights holder or by determining that a title eligible for copyright renewal has not been renewed. For this reason, titles in the public domain do not typically represent current scholarship. Some notable exceptions exist, especially where Hathi has negotiated with scholarly publishers to provide public domain access to recent titles and, to a lesser degree, where individual authors have voluntarily released their claim to copyright on titles in the Hathi archive. Nevertheless, the age distribution for the public

domain content in Hathi is unequivocally skewed toward older titles. ***Approximately 80% of the "full view" books in the HathiTrust Digital Library were published prior to 1923; less than 1% were published in the last decade***. By contrast, if we look at the Hathi corpus as a whole, less than 20% of titles were published before 1923; more than 10% were published since 2000. Clearly, the public domain content represents a relatively mature—not to say more authoritative, or more frequently cited—subset of the scholarly record. It is by no means a representative microcosm.

Similarly, if we consider the distribution of public domain content by topical subject area, it is evident that the scope of coverage differs from that of the HathiTrust Digital Library as a whole. For instance, ***government information constitutes a very small part of the Hathi collection (about 3% of titles in June 2010) but accounts for a disproportionately large share (15%) of titles in the public domain***. By contrast, topical areas that are well-represented in the mass-digitized corpus, and which typically constitute the greatest part of the academic print collection, account for only a very small part of the public domain resource. Titles in language and literature amount to 25% of the HathiTrust Digital Library as a whole, but represent less than 20% of the public domain corpus. Even more remarkable disparities are evident in Art History and Political Science, disciplines where the monograph is a primary vehicle of scholarly communication. Simply put, the "universal library" of digitized public domain content does not represent a microcosm of the academic print collection.

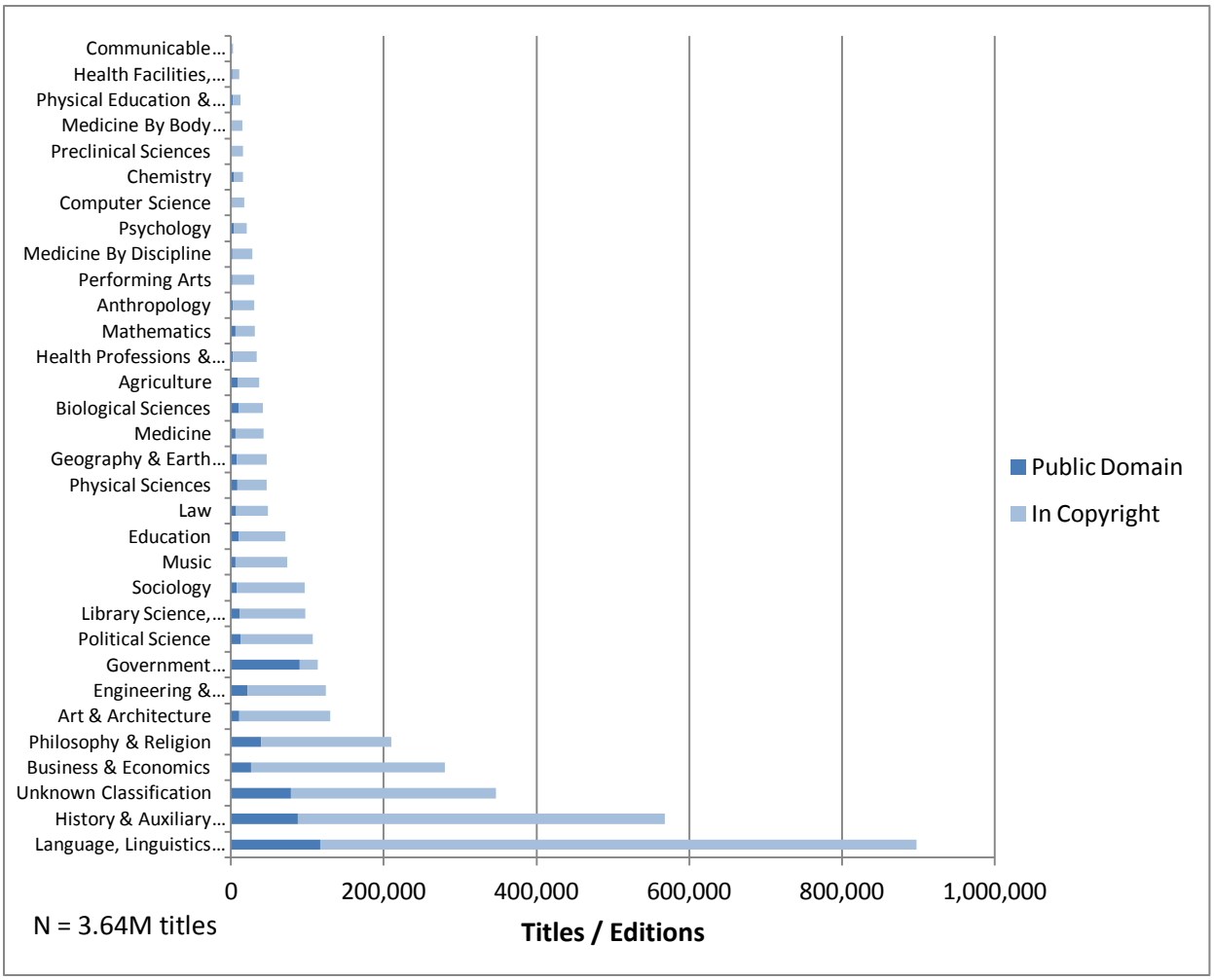


Figure 6. Distribution of titles in HathiTrust Digital Library by subject and copyright status (June 2010)

These findings should not be taken to mean that cooperative agreements aimed at increasing reliance on centralized repositories of digitized public domain content are not worth pursuing. On the contrary, we feel that ***there is substantial opportunity for cost efficient reorganization of academic print collections based on the increased availability of public domain content in the HathiTrust Digital Library***. The sheer magnitude of the HathiTrust Digital Library means that even disciplinary resources that comprise a small proportion of the collection as a whole are, in absolute terms, considerable. For example, Philosophy represents a small fraction of the library (6% of all titles in June 2010), but includes a disproportionate number of titles in the public domain: a total of 39,000, or 19% of all titles in this subject area. Language and literature titles are significantly less likely to be in the public domain (13% in June 2010), but the staggering number of titles in this category means that the net yield—some 116,000 titles—is substantial.

For North American libraries especially, the expanding public domain corpus in the HathiTrust Digital Library represents a shared resource of potentially great value. Although it is unlikely to enable a significant change in local print management operations, it unquestionably improves access to a large body of materials that are otherwise relatively difficult to find or obtain. Because out-of-copyright titles are more likely to represent older and more specialized publications, they are most often held in print by only a small number of academic research libraries with a long collecting history (Lavoie and Dempsey, 2010). As a result, these titles are less visible in the library environment and also more difficult to obtain; their relative scarcity means that they are less likely to be available for inter-lending.

The chart below provides a view of the largest subject-based categories of public domain content in the HathiTrust Digital Library, based on title counts in June 2010. These areas appear to represent the greatest near-term opportunity for redirection of library preservation resources, since at least some libraries can be expected to withdraw and replace locally held physical copies with freely available digital surrogates. At academic and research institutions where off-site and high-density shelving facilities are available, a more systematic and streamlined transfer of low-use print titles from the stacks to storage may be achieved as full-text access eases faculty and librarian concerns about the loss of on-site browsing. Again, the predominance of titles in the humanities is significant, as faculty in history, philosophy and other humanities disciplines are typically the most concerned about relegation of local print inventory. The greater access enabled by full-text provision, in combination with the improved preservation conditions in most off-site facilities, should go some way toward allaying faculty anxiety; if positioned within a larger library strategy for long-term preservation of the scholarly record, it might even embolden faculty to appeal for an accelerated and more aggressive transfer of library holdings off-site.

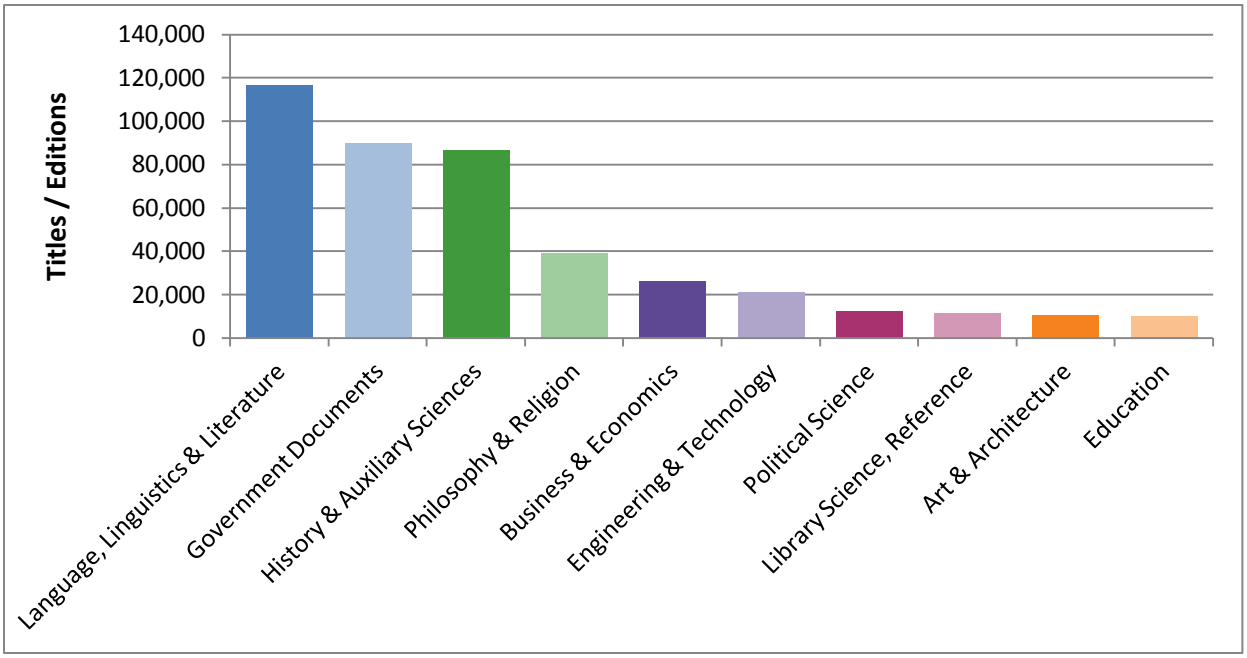


Figure 7. Top ten categories of public domain content in HathiTrust Digital Library (June 2010)

It's reasonable to ask if the current distribution of public domain and in-copyright materials in the HathiTrust Digital Library is likely to change over time, as a secondary effect of an increase in the base of content contributors, in response to a programmatic effort to ramp up public domain contributions or even as a result of the ongoing efforts to renegotiate copyright status. The method we used to harvest and process metadata from the Hathi repository makes it difficult to establish any direct correlation between source of contribution and the relative dearth (or abundance) of public domain content. However, as the proportion of public domain content in academic print collections is relatively low—mirroring patterns in historical print production and library collecting behaviors—even a comprehensive effort to digitize and pool these resources is unlikely to result in a significantly different distribution of public domain and in-copyright titles in the HathiTrust Digital Library. One can reasonably expect that the proportion of "full view" titles and volumes in the shared repository will remain stable at about 16% of titles (20% of volumes) for as long as North American research libraries are the primary source of content contribution.

Distribution of System-wide Print Holdings

The distribution of print holdings for titles in the Hathi repository provides some insights into the potential market for digital preservation and access services. We can predict that libraries will be motivated to redirect management operations (and resources) for print holdings that are replicated in the mass-digitized corpus in proportion to their relative

abundance in the system-wide collection, as well as their rights status and online availability. Simply put, the market value of a digital preservation and access offer that enables many libraries to relegate or withdraw a significant volume of redundant inventory will be greater than the value of a similar offer for titles that are of interest to a smaller number of libraries.

An intriguing and potentially significant finding of our analysis is that many titles in the HathiTrust Digital Library are held by relatively few libraries, based on current WorldCat holdings data. ***Almost 50% of the 3.64 million titles in the repository as of June 2010 are held by fewer than 25 libraries; 14% are held by fewer than 5 libraries***. Put another way, the market for surrogate preservation services for these titles is limited to a small number of libraries who currently own them and who are (in the near term) unlikely to withdraw them, since they represent distinctive institutional assets. The Hathi preservation service offer for these titles would appear to have less (or more accurately, a different kind of) business value, for the specialized audience of research institutions who collectively "care about" the library long tail. A cooperative service agreement shaped around the shared business needs of the ARL community as a whole, rather than the libraries that hold these titles, would possibly provide a means of broadening the base of service and reducing the cost burden for individual Hathi partners. If these relatively rare materials were explicitly marketed as a common-pool resource, cooperatively managed by members of the ARL community, the number of stakeholders prepared to commit resources to Hathi might be enlarged.

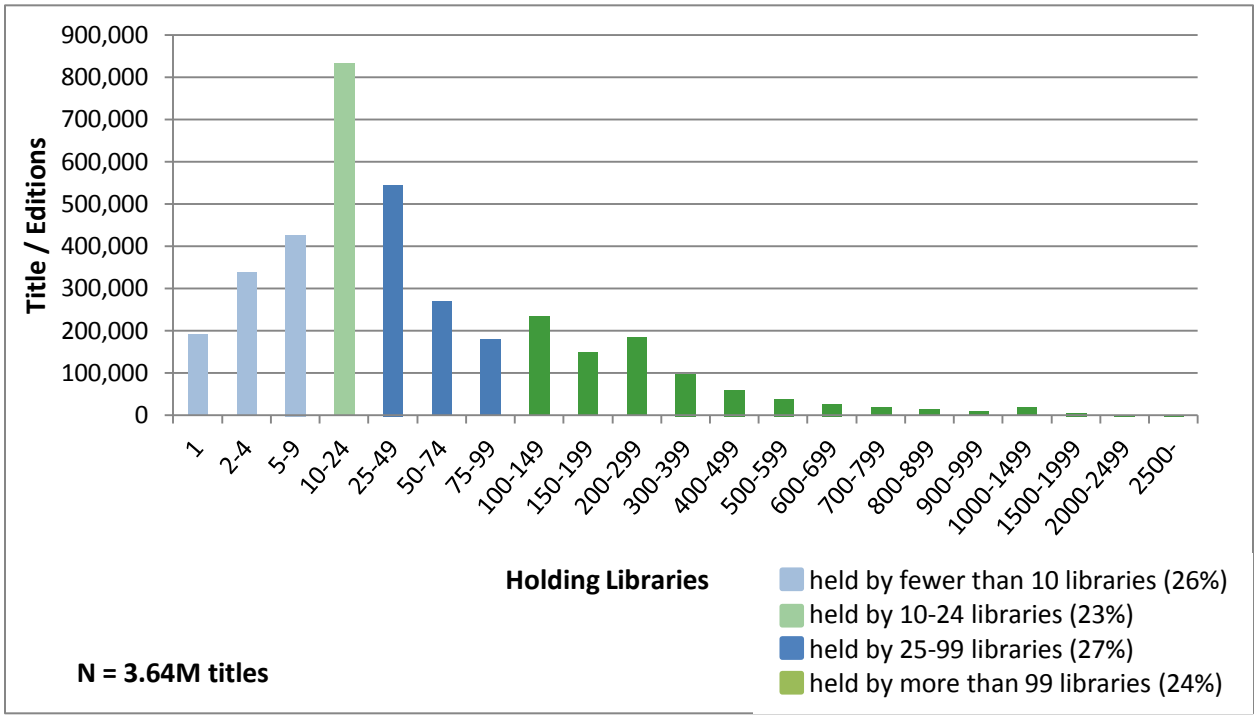


Figure 8. System-wide distribution of library holdings for titles in HathiTrust Digital Library (June 2010)

At the farthest end of the library long tail are ***titles held by a single institution***, for which a redistribution of preservation investment seems most challenging. In June 2010, the HathiTrust Digital Library included more than ***190,000 such titles, representing about 5% of the collection as a whole***. These resources are similar in format and content to uniquely-held print materials examined in previous studies, with an abundance of grey literature, pamphlets, non-English (especially East Asian) titles and, above all, a great number of dissertations and theses (Connaway, O'Neill and Prabha, 2007). Most of the titles in this latter category were contributed by the University of Wisconsin. The rights distribution for Hathi titles with a single holding library is not much different from other titles; approximately 10% are in the public domain. These resources may have great scholarly value, but there is no evidence that they are more accessible as a result of digitization.

The abundance of titles in the HathiTrust Digital Library that are relatively scarcely held should not obscure the fact that ***there is opportunity for significant library space recovery associated with de-duplication of low-use titles*** for which aggregate library supply exceeds projected demand. As of June 2010, there are ***at least 25,000 titles archived in digital format by Hathi for which collective library print holdings per title exceed 1,000 libraries***; more than 900 titles in the HathiTrust Digital Library are held in print by more than 2,500 libraries. It is difficult to imagine a preservation scenario that would require this level

of redundancy in the system-wide print collection. There is considerable debate and discussion in the library community regarding optimal thresholds of duplication in print collections. One widely cited study posits that a minimum of 15 unsecured copies of any given title are needed to ensure survivability of a single copy after one hundred years, assuming typical library loss rates (Schonfeld, 2009). This model presumes an as yet nonexistent network of print preservation guarantees expressed by individual libraries. However, if even a relatively small number of copies are secured in preservation-quality print repositories, a carefully planned strategy to reduce system-wide print inventory is not only theoretically possible but operationally feasible.

As the quality and conditions of use for mass-digitized books continue to improve, as they surely will for titles in the shared Hathi repository, one can imagine that shared print repositories will emerge as an acceptable and even preferred alternative to local management of the mass-digitized book corpus.

Shared Print Repository Profile: ReCAP

A key hypothesis that this study was designed to test is that there is sufficient duplication between shared print storage repositories and the HathiTrust Digital Library to permit a significant number of academic libraries to optimize and reduce total spending on local print management operations. There are at least four library print storage facilities in the United States with holdings in excess of 5 million volume-equivalents that might be supposed to rival the HathiTrust Digital Library in scope of coverage (Payne, 2007). If adequate duplication between these individual repositories and the HathiTrust Digital Library already exists (or can be attained), ***one can imagine a scenario in which client libraries would contract with a regional print repository and with Hathi for preservation and access services, progressively externalizing some portion of local print management operations***. For the purposes of this study, we focused in particular on the Research Collections Access and Preservation consortium (ReCAP) facility, which manages low-use collections deposited by Columbia University, the New York Public Library (NYPL) and Princeton University. In June 2010, the ReCAP collection included more than 8.5 million items.

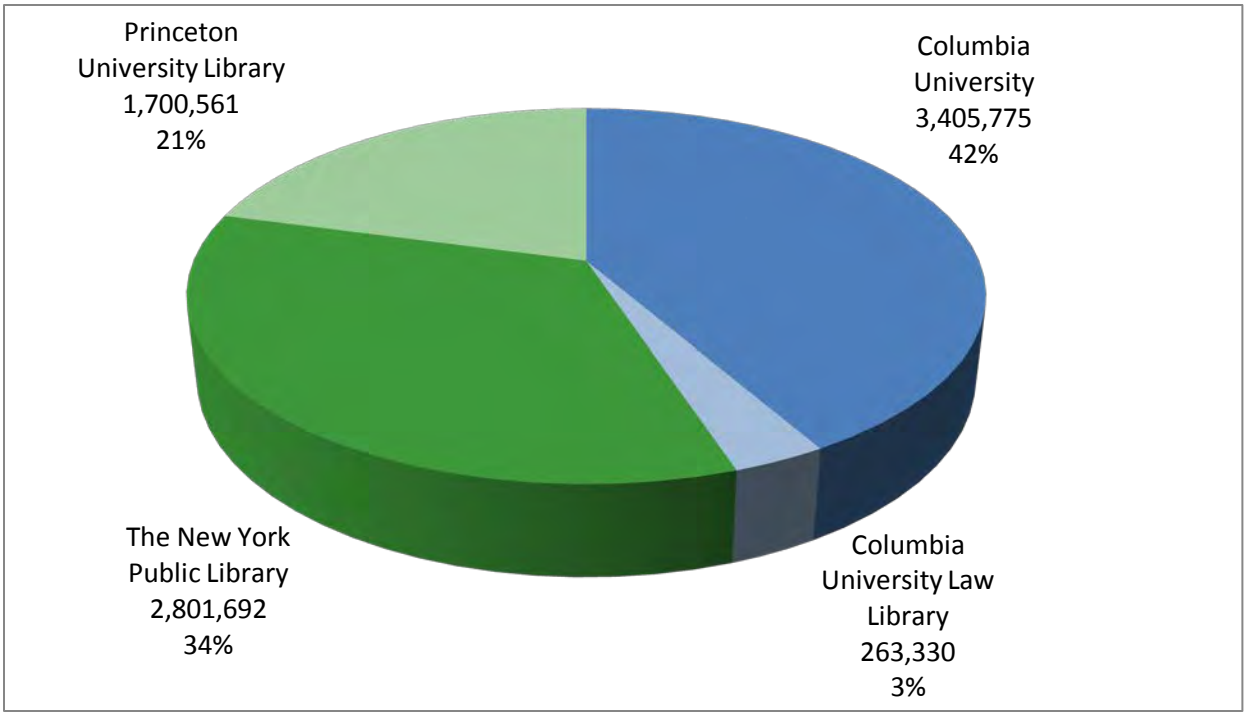


Figure 9. Distribution of ReCAP holdings by contributor (July 2010)

Using sample data provided by Columbia University and NYPL, we examined rates of duplication in ReCAP holdings compared to the HathiTrust Digital Library. Deposits from Columbia and NYPL account for more than 75% of items accessioned by ReCAP, which was considered sufficient for analysis. We were supplied with a sample of approximately four million item-level records (about two million from each library), which were then processed to extract OCLC numbers for matching against the project database. Data from Columbia were processed and merged into the project database in September 2009; data from NYPL were added in March 2010. For this reason, it is not possible to provide a representation of longitudinal changes in coverage of ReCAP holdings replicated in the HathiTrust Digital Library. Moreover, since our ReCAP sample data represents a snapshot of the repository holdings at a discrete point in time, any growth in duplication that we are able to report reflects changes in the composition of the Hathi collection and not new accessions in the ReCAP facility. A further limitation is that because no centralized bibliographic database of ReCAP holdings exists, it is not possible to compare the number of ReCAP titles in Hathi to the number of ReCAP titles as a whole.

Despite these challenges, the data we were able to compile and analyze provide some useful insights. ***Between September 2009 and June 2010, the number of ReCAP titles in our sample that could be matched to titles in the HathiTrust Digital Library more than doubled, from fewer than 300,000 titles to nearly 700,000 titles***. There are a number of

factors contributing to this growth, including some refactoring of code in November which allowed us to map more of the Columbia data to Hathi records, and the addition of the NYPL data in March. It is clear, however, that the rapid pace of growth in the HathiTrust Digital Library also resulted in a net increase in the number of titles that could be matched.

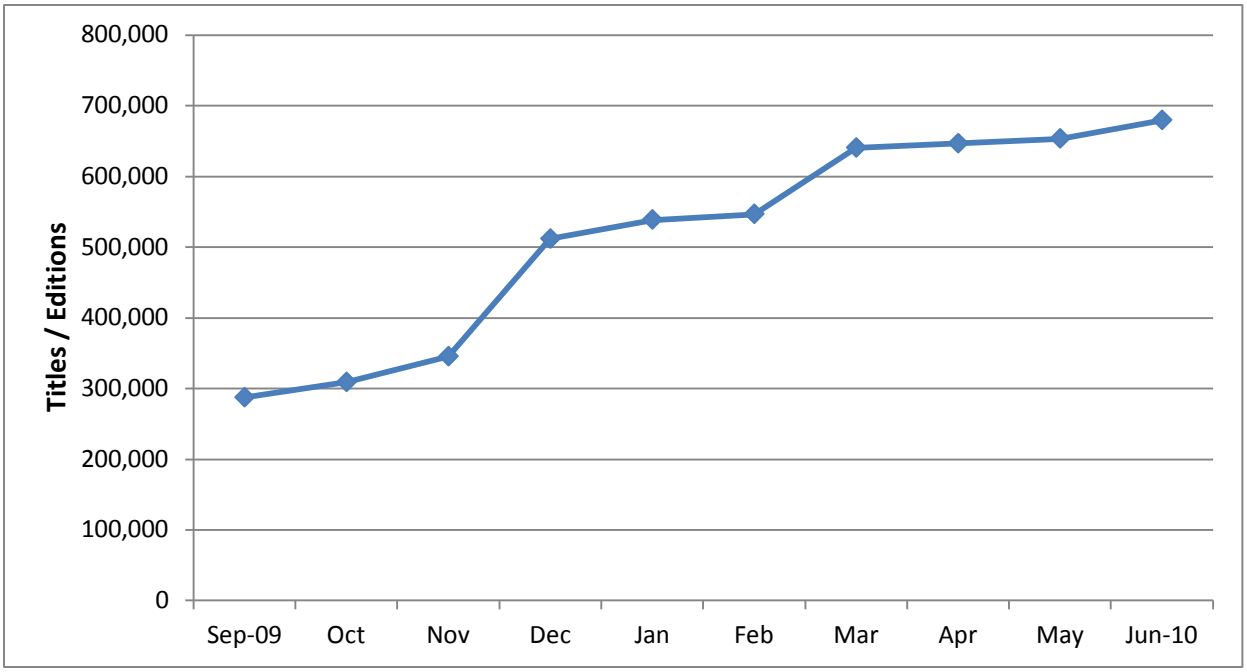


Figure 10. Growth in titles duplicated in ReCAP and HathiTrust Digital Library (September 2009 – June 2010)

Our analysis suggests that ***the ReCAP storage collection mirrors a significant portion of the digitized corpus archived in the HathiTrust Digital Library***; as of June 2010, ***nearly a fifth (19%) of titles preserved in digital format by the HathiTrust are also preserved in print format by ReCAP***. On the surface of things, this may seem like a surprisingly low figure, given our initial premise that the large digital and print preservation repositories were likely to duplicate one another to a large extent. Indeed, we anticipated that the Hathi and ReCAP collections would overlap to a much greater degree, in part because libraries contributing content to the HathiTrust Digital Library were initially drawing on titles digitized from their own offsite storage collections. It seemed reasonable to believe that the digitized collection of titles from storage collections would have a higher probability of being duplicated in ReCAP (or any other large library storage facility) than in an average academic library's circulating collection.

It is possible that a more comprehensive analysis of ReCAP holdings, including titles deposited by Princeton University would result in a somewhat higher Hathi duplication rate. Since Princeton deposits amount to a relatively small part (about 20%) of the total

ReCAP collection, however, it is unlikely that a more comprehensive analysis would result in a substantially different figure. A more probable explanation for the lower than anticipated duplication rate between ReCAP and Hathi is that the scope and character of the large storage repositories from which much of the mass-digitized corpus was initially sourced may differ substantially from the holdings on deposit in ReCAP. Farther below, we explore this thesis by comparing the profile of the ReCAP collection against a few other large-scale depositories.

With these caveats in mind, it is worth considering the potential business value of the ReCAP collection as it mirrors the digitized book collection, on the assumption that an increasing number of academic libraries will seek to externalize print management and preservation in coming years. At the time this project commenced, it was generally believed that the digitized Google Books corpus would be made available as a licensed resource, hastening the trend toward externalization of collection management functions in academic libraries. A year later, the likely outcome of the Google Book Search settlement is still unknown, causing us to question whether university libraries will be motivated to outsource preservation of mass-digitized titles in the absence of a comprehensive licensed access option. Yet ***if the timeline for the digital transition is still uncertain, it is unquestionably the case that academic libraries are being compelled to reconsider the traditional print collection and service portfolio, which was largely dependent on locally managed inventory*** (Michalko, Malpas and Arcolio, 2010). As a strategic reserve, the ReCAP collection and other similar large-scale depositories could thus offer real value even to non-contributing libraries.

In operational terms, the value of a shared print reserve is potentially far greater than traditional inter-lending and reciprocal borrowing arrangements, if shared service agreements for guaranteed access and preservation are in place. For example, an institution like NYU might find it more cost-effective to purchase guaranteed, just-in-case access to print resources managed in a preservation repository than to retain local copies of low-use titles in a legacy collection. ***In the context of a formal service agreement, a library's decision to withdraw local holdings in favor of cooperative preservation and access arrangements would serve a dual purpose of limiting the institution's exposure to risk while reducing the long-term costs of managing local and even remotely stored inventory***.

To understand the degree to which a repository like ReCAP might provide print collection management services scoped around the mass-digitized corpus, it is important to compare not only the relative size of the potential service collection but also its scope and range.

Document Types

As noted above, the emergence of a mass-digitized book corpus presents enormous opportunity for a positive transformation of library service in the academic sector. Substantial operational efficiencies have been achieved in library management of the journal literature as a result of format migration and it is not unreasonable to hope that a similar gain can be achieved for legacy monographic collections. Print book collections are a primary cost driver in academic libraries; while journals occupy a disproportionate share of library space on a per-title basis, the operational expenses associated with acquiring, cataloging and serving monographic collections are substantially higher on a per-unit basis. More pertinently, the long-term carrying costs associated with managing monographic collections have remained largely unchanged. While format migration has enabled many university libraries to shift print journal back-files into more cost-effective storage facilities, low-use print book collections still occupy prime campus real estate, at great expense.

If a shared print service collection is to provide maximum value in the mass-digitized book environment, it is obviously important that it include a very large number of monographs that are also represented in shared digital preservation repositories like Hathi. A potential shared print provider like ReCAP would ideally offer print preservation and access services for a significant number of monographic titles in the mass-digitized corpus and deliberately promote and extend this service collection as a source of distinctive value and utility.

The value of a shared monographic collection of this kind would be different and arguably even greater than that offered by a print journal archive, since uncertainties about the long-term demand trajectory for print books (post-digitization) are likely to sustain a broader and more profitable market for service. Profitability in this context is most likely to be measured in terms of increased efficiency in the academic library enterprise; the marginal gain for cooperative management of books will, at least for a time, be greater than for print journals. This is simply a reflection of the fact that libraries have already made significant strides in lowering the costs of managing the journal literature; the incremental gain that might be achieved by further externalizing journal management is less than is possible (and desirable) for books. For this reason, it is encouraging to find that ReCAP already holds a substantial number of mass-digitized books that could form the kernel of a shared service collection.

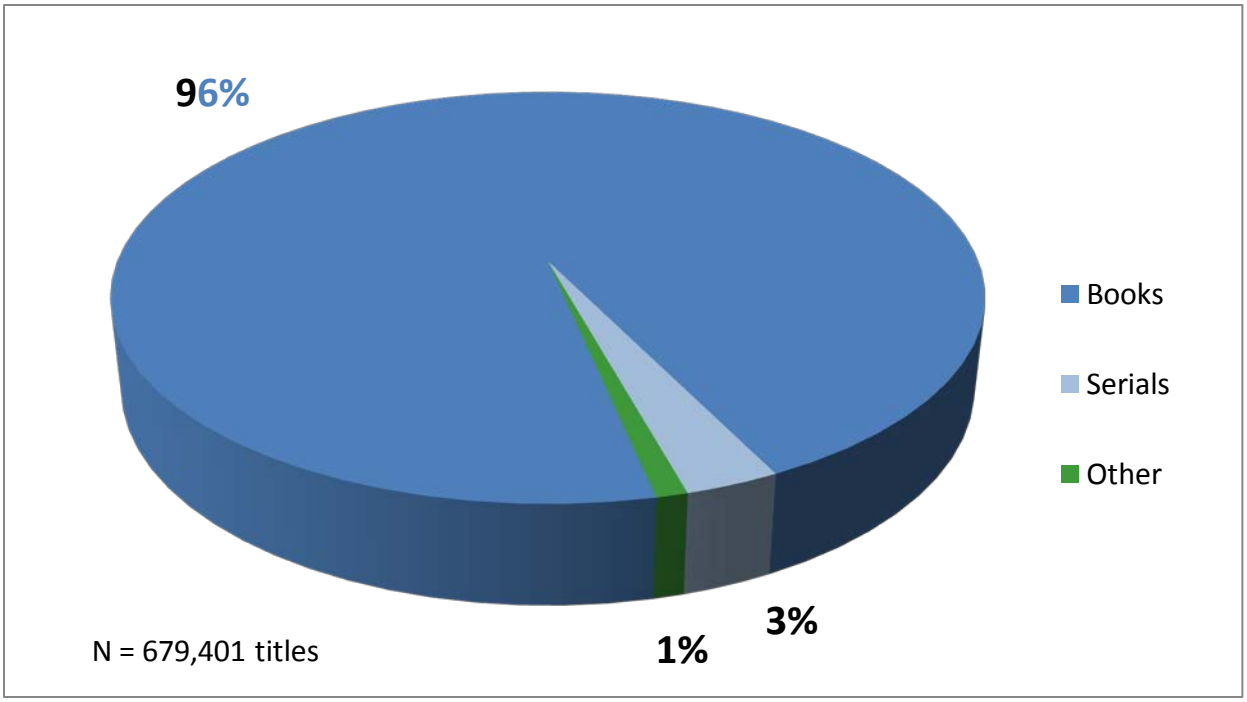


Figure 11. Primary document types of titles duplicated in ReCAP and HathiTrust Digital Library (June 2010)

From a purely pragmatic perspective, implementing shared collection services for a large body of print books may also be somewhat easier than would be the case for serials, where validation of local holdings can be onerous and costly. It is improbable that prospective customers of a shared monographic collection would expect (or pay for) page verification and collation of holdings on a large scale. If required, it could nevertheless be carried out more rapidly and at a lower cost per title for books than for journals.

Subject Distribution

Our examination of the Hathi repository found a preponderance of titles in literature, linguistics, history and other humanities disciplines. We consider this a positive finding, since academic library holdings typically include a large share of humanities titles that occupy a correspondingly large share of the library's physical space. If a significant space savings is to be gained through cooperative management of legacy print collections, it is therefore important that shared service collections include a similarly large share of such titles. Happily, we find that the subject distribution of mass-digitized titles in the ReCAP facility mirrors the distribution of the Hathi corpus as a whole.

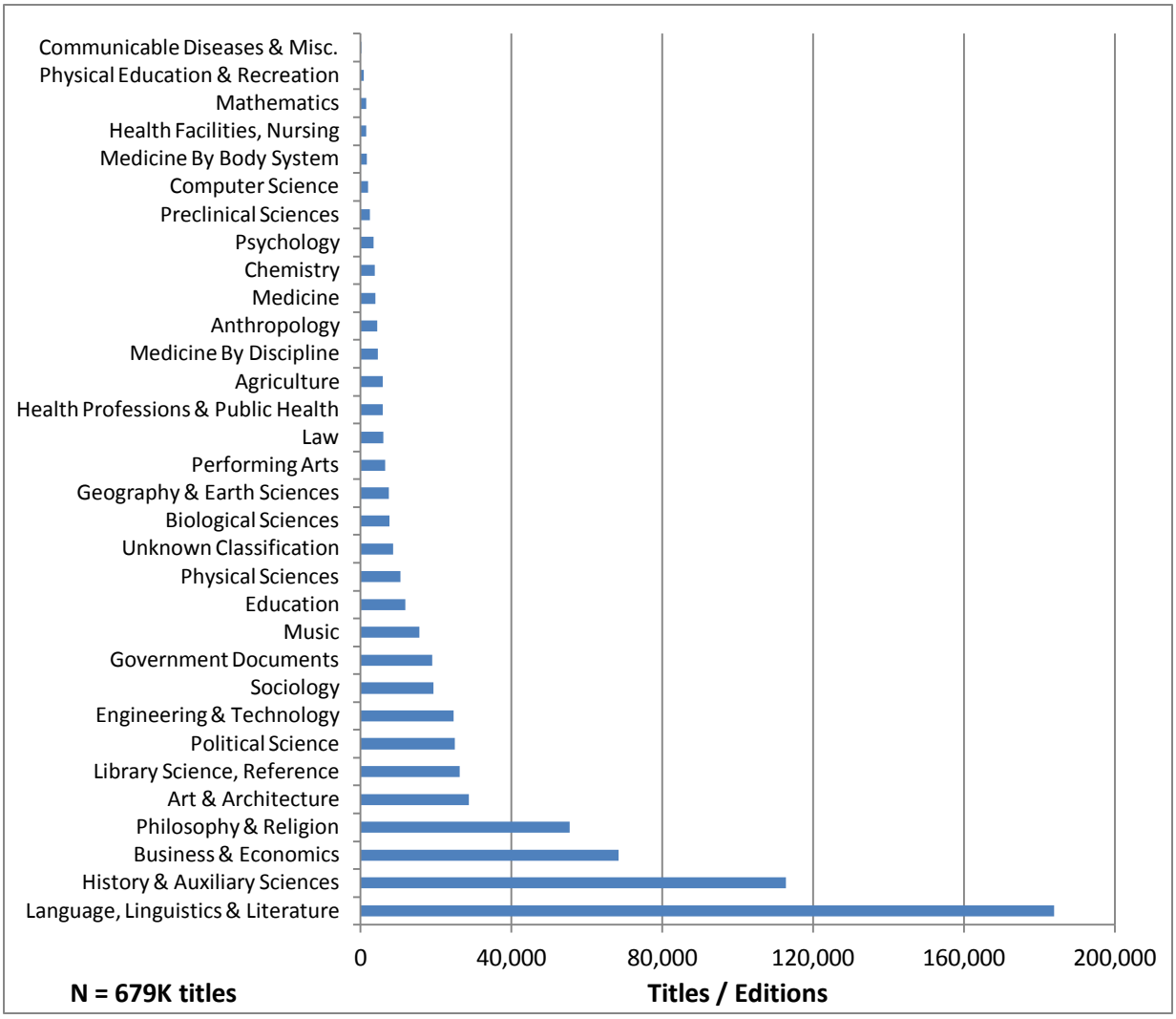


Figure 12. Subject distribution of Hathi titles held in ReCAP (June 2010)

This suggests ***that libraries seeking to "outsource" management of low-use print collections by increasing institutional reliance on shared digital and regional print reserves can realistically expect to transfer preservation and access operations for large monographic collections in the humanities to shared service providers*** like ReCAP, if appropriate service-level expectations are met. It is worth noting that while the disciplinary scope of such an arrangement will be important in building a market for shared services, the business value of the agreements will ultimately be determined by the actual space savings and cost avoidance that can be obtained. A shared print service offer that enables only a modest impact on local operations will likely fail to mobilize sufficient resources to ensure sustainability.

Beyond the extant scope and scale of the prospective shared print collection, additional factors must be considered in evaluating if it is fit for service at scale. The most important of these is its relative availability to an external clientele. This will be determined by both prevailing access provisions and prospective demand.

Rights Status

It is important to assess the relative distribution of public domain and in-copyright content in print preservation repositories like ReCAP, since we can anticipate that demand patterns and preservation expectations are likely to be different for titles that are freely available online and those that are subject to more restrictive authentication regimes. For titles in copyright, especially, it is essential that sufficient stock be maintained on a regional or consortium level, as physical copies will remain an important distribution format for some time to come. Based on the findings of this study, we believe that ***cooperative access and preservation agreements that address the ongoing need for a library print supply chain for in-copyright, digitized books are an essential part of the emerging shared service environment***. Indeed we consider the absence of a collective strategy to build a shared print infrastructure that can meet this need will ultimately expose academic libraries to great risk, as the operational focus shifts away from local management of purchased inventory. Finally, from a purely pragmatic point view, it would appear that shared service provision based on an "insurance only" model, where access to print versions of digitized titles is intentionally restricted to exceptional circumstances—for example, when a print version of a digitized public domain title is expressly required—is unlikely to affect the mobilization of library resource needed to sustain shared print repositories.

A simple illustration will suffice to show that a shared print agreement limited to titles in the public domain can deliver only modest benefit to the academic library community. As noted above, a relatively small part of the mass-digitized corpus in Hathi is available as public domain content. Based on our June 2010 snapshot of the Hathi repository, we estimate that this public domain resource amounts to about 600,000 titles, or approximately 16% of the collection as a whole. ***If ReCAP were to craft a shared print offer around this public domain resource, providing on-demand access to print versions and an assurance of long-term print preservation, it could at best hope to offer a service collection of about 100,000 titles***. A small number of academic research libraries might step forward and commit some ongoing financial support for the long-term care of these materials, even without the assurance that this new investment would be offset by a gain in local operational efficiency. But both the size of the pooled resource and the potential audience for such a service are so small that the total impact would only be marginal to the library enterprise as a whole. ***The economic value of the shared resource in this scenario is further and more***

fundamentally constrained by the fact that titles in the public domain are less widely held than in-copyright titles, so that the efficiencies that might be obtained by consolidating physical holdings as a pooled resource are comparatively slight.

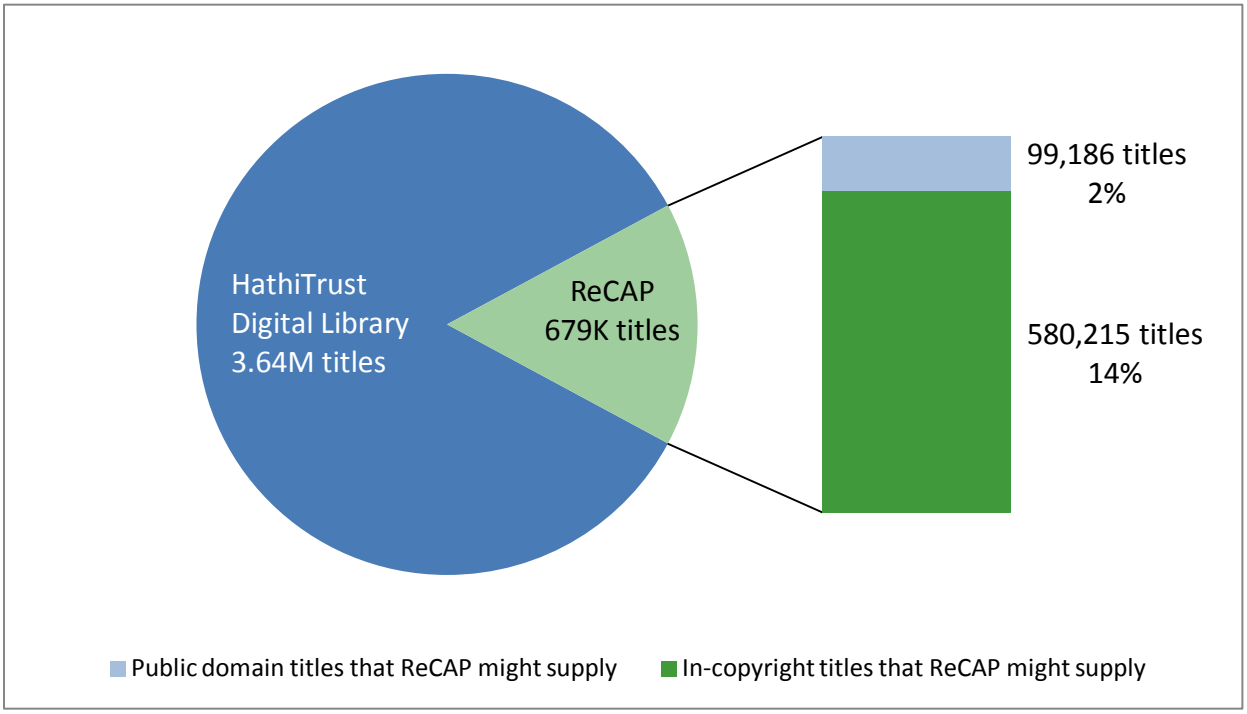


Figure 13. Comparative scope of shared digital and shared print repository collections (June 2010)

By contrast, a ReCAP shared print service offering that included mass-digitized titles currently under copyright could deliver significant value to a large number of academic libraries in the region. In the next section, we consider NYU as an exemplar in the broad market of potential consumers of shared collection services. Here it suffices to observe that as academic libraries look for opportunities to reduce expenditures on operations that can be delivered more effectively at lower cost, it is inevitable that investments in local management of low-use, legacy print collections will come under close and critical scrutiny. ***Even, and perhaps especially, in advance of an eventual licensing agreement that will enable libraries to retire local print collections in favor of digital aggregations—a transition that is both deeply feared and fervently desired by scholars and librarians alike—it is imperative that academic administrators begin to plan for a library future in which management operations are selectively and strategically shifted outside the local institution, to larger regional and consortial interests***. Unless this transition is proactively managed by library directors and supported by the academic institutions they serve, there is a strong probability (not to say certainty) that the legacy print collections we have long cultivated as

institutional assets will eventually be regarded as local liabilities. Unless the collective value of these resources is accounted for and memorialized in new sets of inter-institutional agreements, responsibility for the preservation of these resources will likely devolve to a handful of research institutions not adequately equipped or empowered to assume a "permanent" stewardship role.

Availability of Repository Holdings

Materials on deposit in the ReCAP repository are subject to a variety of access rules imposed by the depositing libraries and library units. Thus, books deposited by the Avery Architectural and Fine Arts Library at Columbia University may be subject to circulation and lending restrictions that make them less available than materials deposited by the main Nicholas Murray Butler Library. We used location codes harvested from the ReCAP sample data to examine the relative availability of print versions of titles in the HathiTrust Digital Library, with the expectation that ***availability of repository collections is certain to be a key factor in negotiating shared print agreements with noncontributing partner libraries***. Just as the availability of digitized content in the HathiTrust Digital Library is constrained by intellectual property rights enshrined in copyright law, the availability of print repository holdings is constrained by access rules imposed by owning libraries. Understanding the scope of these constraints is essential to assessing the feasibility of a truly scalable approach to shared service provision.

Fourteen different location codes were included the ReCAP sample data we analyzed; thirteen are associated with Columbia University campus libraries. The chart below reveals the distribution by location or "customer code" of titles in the ReCAP sample that could be matched to digitized titles in the HathiTrust Digital Library in June 2010. Note that because there is some duplication in collections on deposit in the ReCAP collection, the number of ReCAP holdings replicated in Hathi (714,955 volumes) is greater than the number of titles (679,401) that are held in common by both ReCAP and Hathi. As shown below, ReCAP deposits from Butler Library (CU Standard) and NYPL account for the majority of holdings. This is a positive finding, since ReCAP holdings from these libraries are largely unrestricted and therefore potentially in scope for a shared print service agreement. ***As of June 2010, the ReCAP collection included nearly 600,000 unrestricted titles that mirror content in the HathiTrust Digital Library; this represents a significant pool of resources that might be marketed as a shared print collection***. Put another way, almost 90% of the ReCAP collection that is potentially in scope as a surrogate service collection for mass-digitized content could be transitioned into a shared service model without disrupting the current accessioning model.

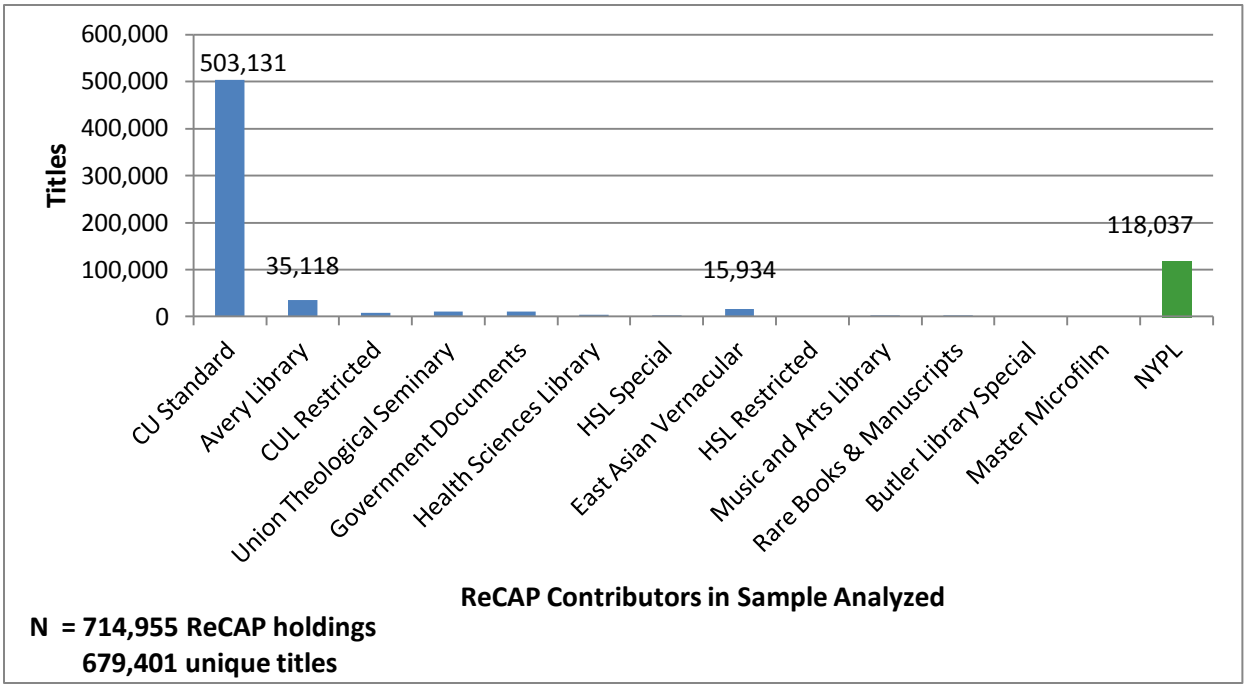


Figure 14. Titles duplicated in ReCAP and the HathiTrust Digital Library (June 2010)

Potential disruption to the current ReCAP business model is an important consideration in assessing the costs and benefits of a transition to a shared print service model. However, it is equally ***important to consider where changes in current operations might significantly improve ReCAP's ability to serve as a shared print service provider***. For instance, a simplification and normalization of circulation rules associated with the 70+ customer codes would enable external clients to more easily evaluate the business value of a partnership with ReCAP. A systematic effort to consolidate ReCAP holdings under a few common access regimes would maximize the business value of the repository and likely result in more cost-effective management of the pooled resource. One obvious benefit of a cooperative collection management regime would be the space savings obtained by de-duplication of holdings transferred to the shared facility. Beyond extending the useful life of the current facility, a selective deduplication effort would also increase the potential scope of a shared print service offer by increasing the range of titles represented in the resource.

Our study identified ***more than 25,000 Hathi titles deposited in ReCAP by both Columbia University Libraries and NYPL***. Using a cost estimate of $.86 per volume for management in a high-density facility, one can estimate that the ***ReCAP consortium is investing at least $40,000 annually in the management of print inventory that is duplicated within both the shared repository (with multiple partner copies on deposit) and the Hathi digital preservation repository***. The long-term cost of preserving these resources could be reduced

by half or more if duplicate copies were removed from the repository or, better still, not accessioned into the repository. This is a conservative estimate, based on the number of monographic titles duplicated in both ReCAP and Hathi for which duplicate deposits by NYPL and Columbia could be identified. The actual cost figure is likely to be much greater, since in some cases multiple copies of a title have been deposited by each partner library. For example, Columbia University may deposit several copies of the same book, each under a different departmental customer code. We identified tens of thousands of titles in ReCAP deposited by multiple departmental libraries at Columbia University, which are also replicated in the HathiTrust Digital Library.

There are undoubtedly instances where duplication in the aggregate ReCAP collection is justified; for instance, when a title is rare or of special cultural or institutional significance. Our findings suggest that many titles duplicated within the ReCAP repository (i.e., deposited under multiple customer codes) are also held by hundreds of other libraries. As might be expected, titles deposited by multiple ReCAP partners are generally more widely held in the system-wide library collection than titles deposited by a single ReCAP partner. About a third of Hathi titles in ReCAP that are on deposit by a single library partner could be described as relatively widely-held titles, with more than 99 library holdings in the WorldCat database; nearly 50% of the titles deposited by multiple partner libraries fall into this "widely-held" category. Under the present arrangement, the collective investment made by ReCAP to manage these duplicate copies represents a significant opportunity cost: every dollar spent to store a second or third copy of a widely held book is a dollar that can't be spent on a potentially higher priority item.

Strictly speaking, it may not be possible for ReCAP to reduce significantly its expenditure on managing redundant inventory already accessioned in the repository. Given the effort and expense required to deduplicate and restock inventory in a high-density repository, it seems unlikely that a retrospective de-selection of ReCAP holdings will be undertaken. The consortium could, however, ***maximize the value of its ongoing investment in the repository collection by making it available to external library partners as a shared preservation resource***, analogous in some respects to the shared Hathi digital repository. By deliberately accessioning materials for which a broad market for service exists, and by managing the pooled inventory as a cooperative resource, ReCAP partner libraries can substantially increase the business value of the shared repository.

Distribution of System-wide Print Holdings

Finally, to understand the potential value that a shared service offering from ReCAP might deliver, it is useful to consider the market it would likely serve. Since libraries that hold local copies of titles duplicated in ReCAP and Hathi have a shared interest in the long-term

preservation of these resources and a collective interest in reducing unnecessary expenditure, we can look to the distribution of library holdings in ReCAP's prospective service collection for an indication of its potential market for service.

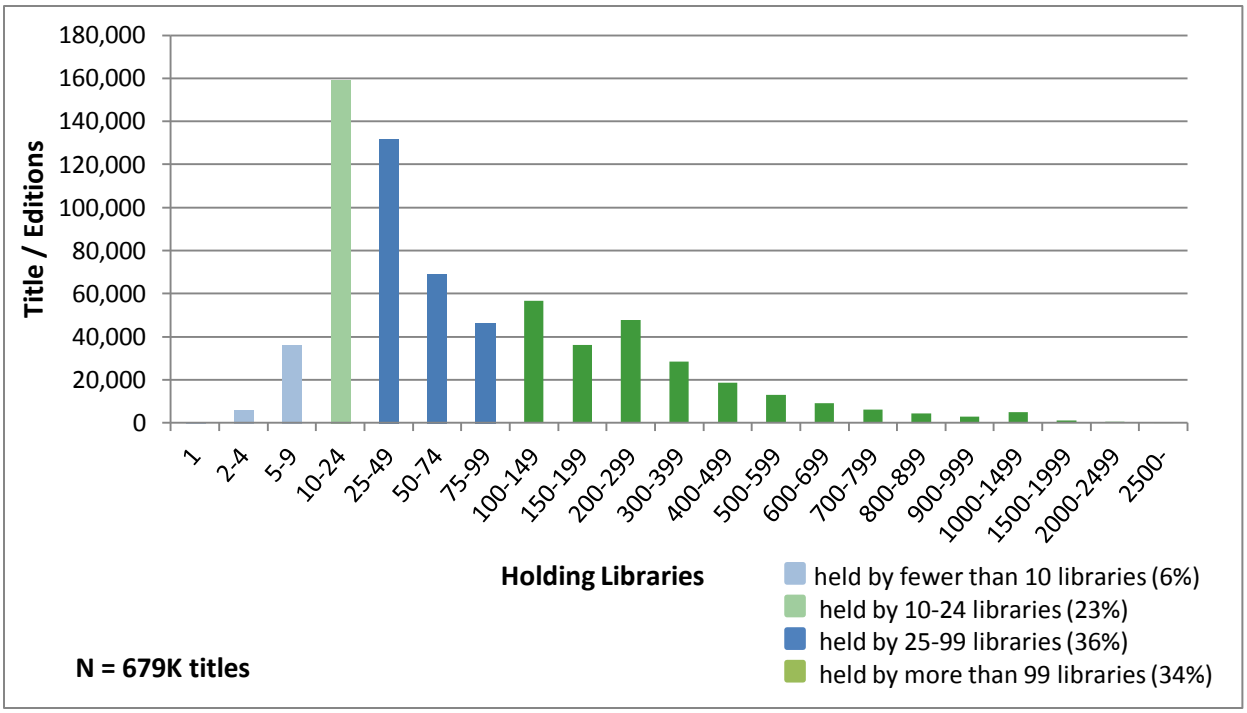


Figure 15. System-wide distribution of library holdings for Hathi titles in ReCAP (June 2010)

A relatively small proportion of the titles duplicated in ReCAP and Hathi represent rare or scarcely-held resources. Less than 10% of the titles we examined are held by fewer than 10 libraries. From a shared print service perspective, this is a positive finding since we can predict that relatively few libraries are prepared to allocate significant resources to ensure print preservation of these materials, unless they can be shown to have special cultural or scholarly importance. In a shared repository environment, the costs of preserving and providing access to these resources will be comparatively low, but the value they represent to potential consumer libraries is also relatively small. By contrast, titles that are held by a larger number of libraries are likely to have a greater business value; every library that acquired this content has a greater or lesser stake in its long-term preservation and many (if not most) of them will be motivated to reassess the costs and benefits of local management once the title is represented in the mass-digitized corpus.

Based on our analysis, ***the ReCAP repository holds more than half a million mass-digitized titles that are also held by 25 or more libraries***. This inventory could have considerable business value as a shared print service collection. Compared to the distribution of library

holdings for titles in Hathi, the ReCAP profile shows a greater concentration of widely-held books, with 70% of titles held by more than 25 libraries. This is not an exceptional finding (there is a greater probability that ReCAP will hold a title that is relatively abundant in the larger library system versus a title that is rare) but is an important one. It suggests that ReCAP could establish a market as a shared print provider for a substantial number of libraries. This has obvious implications for future business planning.

Model Consumer Profile: NYU

Analysis of the duplication between NYU library holdings and the Hathi repository has confirmed a hypothesis framed at the outset of this project: the emerging corpus of digitized books represents a potentially viable surrogate for a substantial proportion of print book collections in academic libraries, if adequately "backed up" or reinforced by a shared print access and preservation strategy. ***In June 2009, approximately 20% of titles in NYU's Bobst library*** (as measured by holdings in WorldCat) ***were duplicated in the Hathi repository; by June 2010, the rate of duplication had increased to about 30%***. It is tantalizing to consider the space recovery and cost avoidance that might be achieved if the library could outsource preservation and access services for at least some of these titles to shared print and digital repositories.

In absolute numbers, the overlap in titles held by NYU and Hathi is significant. Our June 2010 analysis identified more than ***700,000 titles*** (unique editions or manifestations) that were ***held in both repositories***, i.e., archived in digital format by the HathiTrust and held in a tangible (usually print) format by NYU Libraries. This constitutes almost a third of the Bobst library collection, on a per title basis. Based on standard volume-equivalent measures, it represents approximately ***44,000 linear feet of standard library shelving*** or about 55,000 assignable square feet (ASF) that could be repurposed for new uses. The chart below documents the growth in duplication between NYU and Hathi holdings over the twelve months of our study.

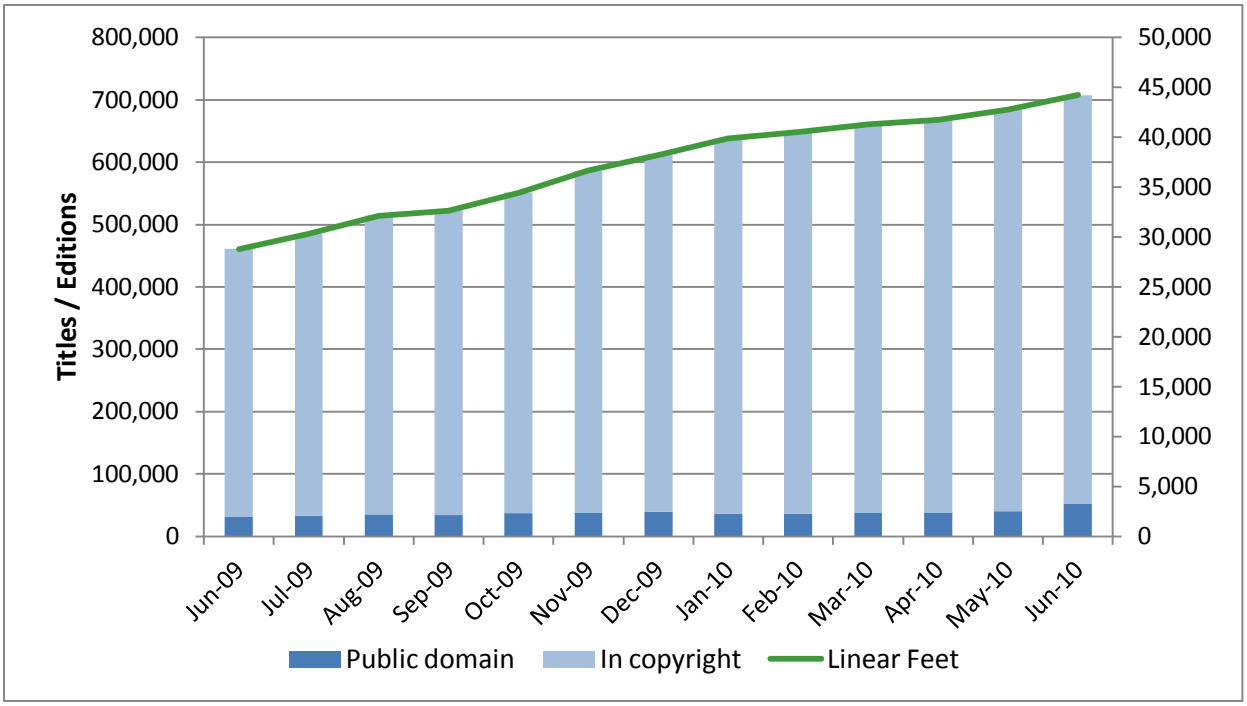


Figure 16. Growth in coverage of NYU Bobst holdings in HathiTrust Digital Library (June 2009-June 2010)

In addition to space recovery, the potential cost avoidance that might be achieved if redundant inventory were permanently removed from the library is considerable. Using a recently published cost model produced by economist Paul Courant, one could calculate the ***total annual cost savings that might be achieved by replacing 700,000 locally managed print volumes with a surrogate service provision to be as much as $3 million per year***—assuming (somewhat improbably) that all staffing, operational and facilities expenditures were adjusted to reflect the change in collection size (Courant and Nielson, 2010). Yet if even a small fraction of this cost avoidance could be achieved, it is obvious that the library might substantially reduce or, more strategically, redirect its draw on the financial resources of the university.

In reality, of course, it is unlikely that NYU or any other research institution would view duplication of even very low-use local physical holdings with titles in the Hathi archive as sufficient justification for permanent withdrawal. As noted above, an overwhelming majority of titles in the digital archive are still in copyright and therefore subject to restrictions in online availability; NYU cannot simply "replace" access to locally held physical inventory with a link to a free digital edition. Online access to these titles will ultimately require a subscription to Google Books or another licensed aggregation. More importantly, local faculty—especially humanities scholars—will almost certainly expect NYU to provide ongoing access to print versions of titles that were purchased by the library, irrespective of online

availability. This leads us to the question of whether a regional storage collection like ReCAP can provide a more cost-effective solution to long term physical preservation and access, for titles that have been digitized and securely archived by Hathi.

The level of duplication between NYU, Hathi and the ReCAP facility provides a baseline measure of the potential savings that might be achieved in the near term.

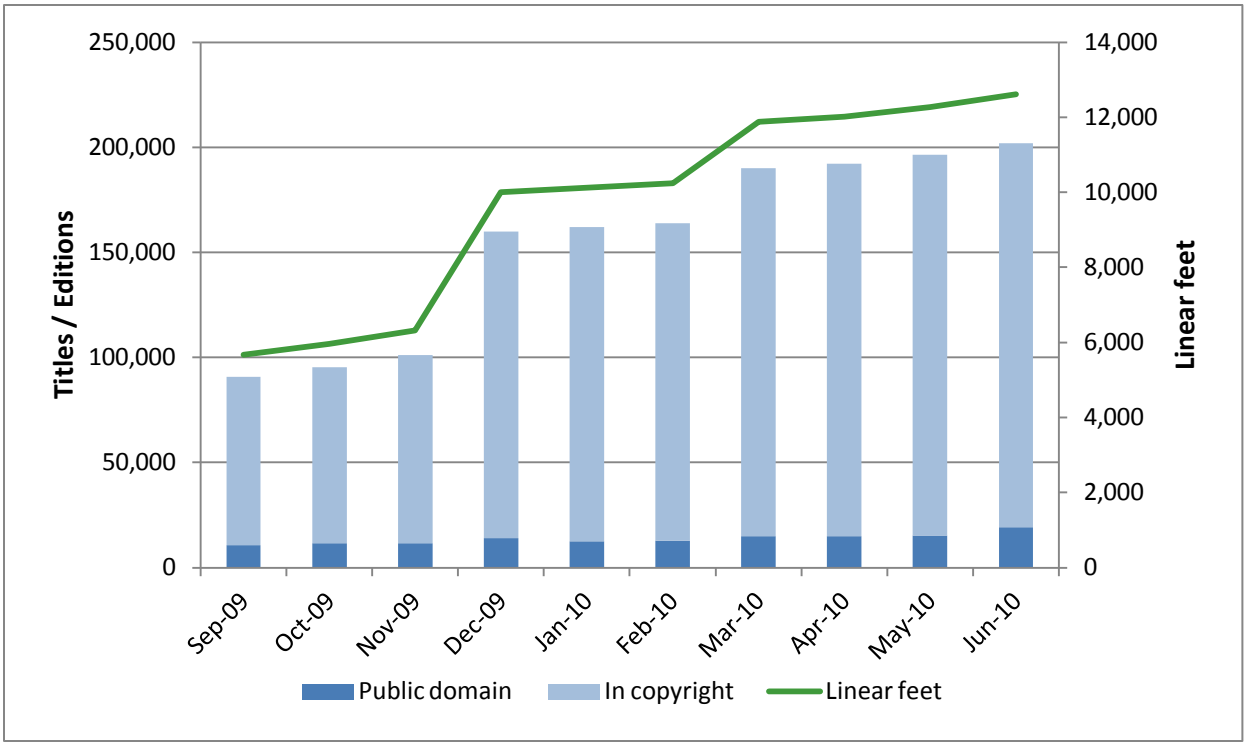


Figure 17. NYU Bobst titles duplicated in ReCAP and HathiTrust Digital Library (September 2009–June 2010)

A relatively small proportion of the titles owned by NYU that are currently replicated in the HathiTrust Digital Library were found to be duplicated in the ReCAP sample data. ***Based on our June 2010 analysis, we identified about 200,000 titles that might be eligible for withdrawal from NYU's Bobst Library collection, based on "dual duplication" in Hathi and ReCAP, if a robust shared service agreement were in place***. This figure represents about 10% of the Bobst Library's total holdings in WorldCat, a relatively modest figure compared to the 30% of Bobst holdings that are replicated in the mass-digitized corpus in Hathi. If we limit the analysis to monographic titles in the public domain, the total number of books that would be considered in scope for a shared print agreement between NYU and ReCAP dwindles sharply, to about 18,000. In short, the

initial premise that a cooperative management regime restricted to books in the public domain might deliver sufficient benefit to mobilize a significant change in local library operations is seemingly not borne out by our findings.

If we look instead at the full range of Bobst holdings in Hathi—including in-copyright titles—ReCAP appears to be a potentially more valuable supplier of print preservation and access services. More than 90% of the 200,000 titles in Bobst that are duplicated in Hathi represent in-copyright content; nearly all of them (93%) are designated as unrestricted holdings in the ReCAP collections. A print supply chain for these titles will remain indispensible for some time to come and whether post-digitization demand increases or falls, centralized repository services will prove more cost-effective solution than local management. Accordingly, one might concede that ***a shared print service offer from ReCAP that includes any mass-digitized title owned by NYU would maximize the library's ability to reduce unnecessary expenditure and redeploy resources in support of services*** more directly tied to the university's research and teaching mission.

Even so, it is not clear that a shared print service arrangement with a sole provider like ReCAP can deliver the level of benefit that is ultimately desired. Is a shared print offer that enables a net space recovery of less than 13,000 linear feet of shelving (or approximately 16,000 ASF) in its first year worth the energy and effort that would be needed to draft and implement a binding agreement? At the time this study began, NYU was in the midst of a major library renovation project that required the removal of nearly 500,000 volumes from the Bobst library building. Under normal operating circumstances, the library routinely transfers 100,000 volumes or more to storage each year, simply to accommodate annual growth in the print collection. By comparison, the potential space savings and cost avoidance achieved by outsourcing collection services for 200,000 books in ReCAP appears relatively small.

Realistically, one can estimate that a potential consumer of shared print services might hope to "externalize" or outsource management for all, or almost all, of the mass-digitized content in the local collection. Thus, NYU might reasonably seek to negotiate a shared service agreement that would cover most of the 700,000 locally owned titles that are currently duplicated in the HathiTrust Digital Library, while also allowing for additional growth in years to come. Under present conditions, a shared print agreement with a single supplier with ReCAP's current collection profile would deliver less than a third of the value that NYU might hope to derive from a fully satisfactory shared print agreement that provided comprehensive coverage of the mass-digitized corpus.

Does this mean that academic libraries must postpone any space and cost saving reorganization of low-use print collections until an eventual e-licensing option for the mass-

digitized book corpus emerges? We think not. A range of options are available to individual academic institutions seeking to externalize some part of the operational and cost burdens associated with managing a low- or no-use legacy print collection. Several of these options are explored below.

Shared Print Provision: Assessing the Options

As outlined above, our analysis of the NYU library holdings duplicated in the HathiTrust Digital Library and ReCAP suggests that the total space savings and cost avoidance that is achievable through a shared print agreement as originally conceived is relatively limited at present. This outcome is dependent on a number of variables, at least some of which are subject to library influence or control. Libraries seriously motivated to design and implement shared service agreements will want to consider all available options for maximizing the impact and sustainability of these models.

Expanding the Scope of Shared Service

In examining the NYU library collection, we limited our study to holdings in the Elmer Holmes Bobst library, which serves the general undergraduate population as well as graduate researchers and faculty in the humanities, social and natural sciences. Bobst is the largest of NYU's libraries, with holdings in excess of 5 million volumes, and ranks among the top academic research libraries in the United States. Our analysis found only a modest overlap between holdings in Bobst, Hathi and ReCAP, amounting to less than 10% of all titles in the Bobst library collection. By contrast, the duplication rate between Bobst and Hathi alone is estimated to exceed 30% of the local collection.

Based on library holdings registered in the WorldCat database, we estimate that holdings in Bobst account for about 85% of all titles held by NYU libraries. Conceivably, a greater yield might be obtained in a shared print agreement with ReCAP if the scope of our analysis were adjusted to include a wider range of NYU library units. To test this hypothesis, we expanded the scope of comparison to include all of the NYU libraries with holdings set in WorldCat and identified a total of 775,980 unique titles replicated in Hathi as of June 2010. This represents a 10% increase over the number of Bobst titles in Hathi (679,401 titles). Yet while the base of comparison for a shared print offer for NYU was larger in absolute terms, the proportion of titles that ReCAP might supply actually decreased to 28%, compared to 29% for Bobst alone. In a very real sense, one can say that expanding the scope of an initial shared print service offer to include a broader range of library types (including specialized departmental libraries) would actually result in a lower net benefit. Expanding the scope of a shared print arrangement would also entail more complex business agreement since not all NYU libraries are under the same administrative or budgetary control.

This leads us to conclude that an initial shared print offer should in the first instance be scoped around the space- and cost-saving objectives of a limited range of academic libraries that share a common set of service expectations. Based on what we have seen from the shared digital and shared print repository profiles, the target audience is likely to be moderate to very large college and university collections in the humanities and social sciences. ***One can predict that mid-size universities with a strong commitment to the humanities, along with liberal arts colleges, will represent the core market for shared monographic preservation and access services, since they are committed to uphold a preservation mandate for which local resources are increasingly inadequate.***

Assessing Market Maturity

The findings reported here are based on a twelve-month study of the mass-digitized corpus in Hathi and a single, partial snapshot of the ReCAP print repository. As shown in figures 10 and 17 above, the prospective value of ReCAP as a shared print service provider has increased significantly in the past year based on the rapidly expanding scope of Hathi alone. During this same period, the ReCAP collection itself has also grown, with about 50,000 new items representing an unknown number of unique titles accessioned each month on average. Is it possible that the "dual duplication" rate that was predicated as necessary for shared service provision—with at least one copy in Hathi and one copy in a shared print preservation repository—will increase as the ReCAP inventory continues to grow? If so, is it worth waiting until the duplication between ReCAP and Hathi reaches a desired threshold? This question is especially pertinent in the case of ReCAP since all three ReCAP partner libraries have joined the HathiTrust since the inception of this project. These libraries are now more likely to transfer to ReCAP the titles digitized from their local collections, which will naturally increase the match rate between Hathi and ReCAP.

The evidence in hand suggests that unless a deliberate and systematic effort is made to align shared print repository holdings and Hathi digital repository holdings, it is unlikely that the existing preservation infrastructure embodied in large library storage collections will coalesce into a sufficiently robust source of surrogate supply. What is required is not an incremental and ad hoc change in storage transfer protocols at individual repositories, but a purposeful and coordinated strategy to create a shared print infrastructure capable of delivering significant tangible benefit to a large number of academic libraries. Deferring the negotiation of shared print agreements until such time existing repositories exhibit the desired service profile will simply delay the development of shared infrastructure. Instead***, library administrators who readily perceive and are prepared to realize the benefits of selectively outsourcing print management functions to a shared service provider can accelerate the process by stipulating clear expectations and establishing targets that prospective providers will be motivated to meet.***

Alternative Service Providers

We have seen from our case study that the current ReCAP repository collection, as represented in our sample, is not optimized for shared print provision as originally envisioned. At best, it might at the outset offer a client such as NYU surrogate preservation and access services for about a fifth of the mass-digitized book collection in Hathi and enable a local space recovery of about 13,000 linear feet, or the equivalent of about 200,000 volumes. Even if this figure were to grow in future years, as will almost certainly happen, the initial value proposition appears insufficient to justify a service agreement, except perhaps as a symbolic gesture.

Could another large-scale print repository provide a more competitive offer? The Southern Regional Library Facility (SRLF), a shared library storage facility serving five campuses in the University of California (UC) system, currently holds about 6 million items, a collection about three-quarters the size of ReCAP. The SRLF provides a useful counterpoint to ReCAP, since it represents holdings from a more uniform base of academic institutions and encompasses a broader range of university libraries, including both ARL and non-ARL institutions. Inventory at the SRLF is managed as a cooperative resource with a non-duplication policy applied across the aggregate collection. Thus, one might expect that the SRLF collection would be more broadly representative of academic library holdings and also contain a greater proportion of unique titles than is the case at ReCAP. The SRLF is remarkable in another way as well: it is one of the few large library storage collections whose holdings are represented in the WorldCat database under a discrete library symbol.

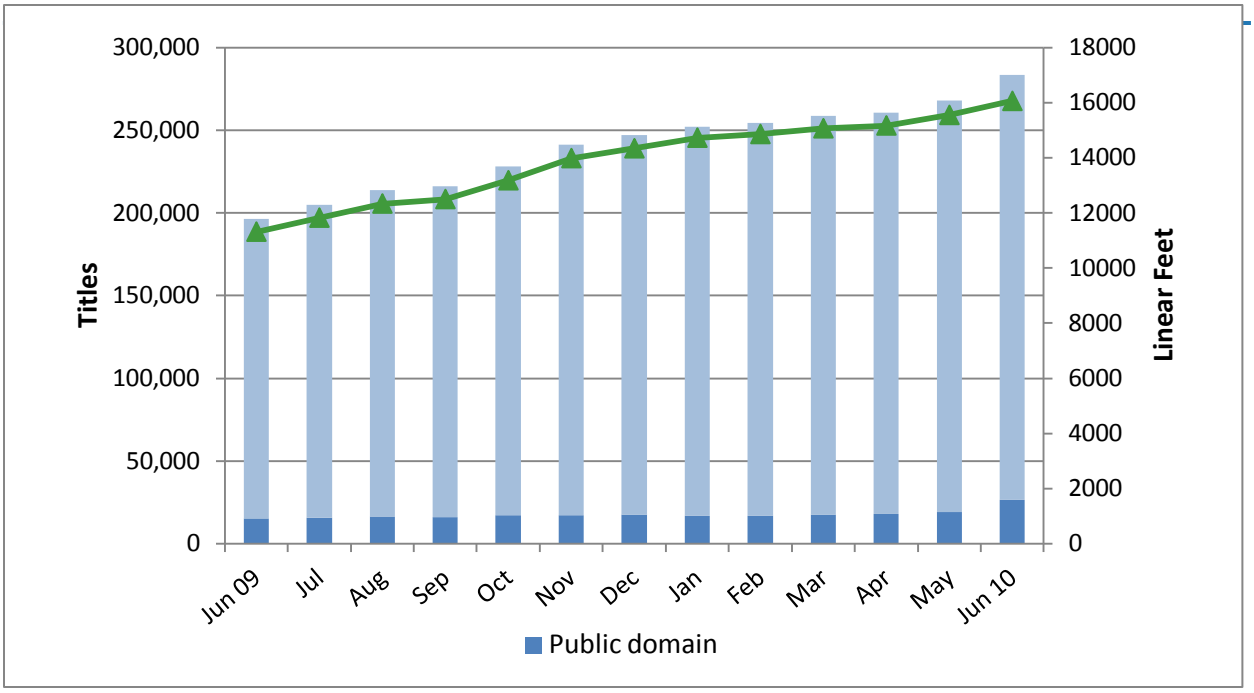


Figure 18. NYU Bobst titles duplicated in UC SRLF and HathiTrust Digital Library (June 2009–June 2010)

Although smaller in absolute size, the SRLF appears to offer greater initial value as a shared print partner for NYU. Comparing figures 17 and 18, one can see that the SRLF would enable a marginally greater space savings and potential cost avoidance than ReCAP. More significantly, perhaps, it appears that the rate of growth in coverage over time is sustained and relatively stable—important factors when negotiating a business agreement that is in part predicated on a forecast of future value. Additionally, a visible spike in the proportion of public domain titles that the SRLF could supply between May and June 2010 serves as a useful reminder that a shared print repository that proactively contributes to the development of a shared digital infrastructure simultaneously increases its value as provider of surrogate print services. The increase in public domain titles held in common by SRLF and NYU in June 2010 reflects a large Hathi ingest of titles digitized by the University of California in partnership with Microsoft and the Internet Archive.

The UC SRLF example is instructive, but not necessarily representative of the broader marketplace of potential shared print providers. Over the course of this project, another large library preservation repository fortuitously became visible in WorldCat: the UC Northern Regional Library Facility (NRLF), which serves five campuses in Northern California. While approximately the same size as the SRLF, with approximately 5.5 million items held in June 2010, the NRLF offers significantly less coverage of the mass-digitized book corpus held by NYU, though still more than ReCAP can presently provide. This is due in part to the fact that the NRLF holds a greater proportion of rare and unique titles than its Southern California

counterpart; as a result, there is a lower probability that titles in the repository will be duplicated in Hathi (unless directly contributed as digitized content) or in other library print collections. As noted above, repositories that hold a greater relative proportion of titles with very low aggregate holdings will probably find it difficult to establish significant market share in a shared print service environment.

Compared to ReCAP, the University of California's two massive regional repositories offer only slightly greater coverage: the UC NRLF might provide a provisioning option for 30% of the titles of interest (compared to 29% at ReCAP); the UC SRLF could potentially provide preservation and access services for 36% of the assumed target of 700,000 volumes.

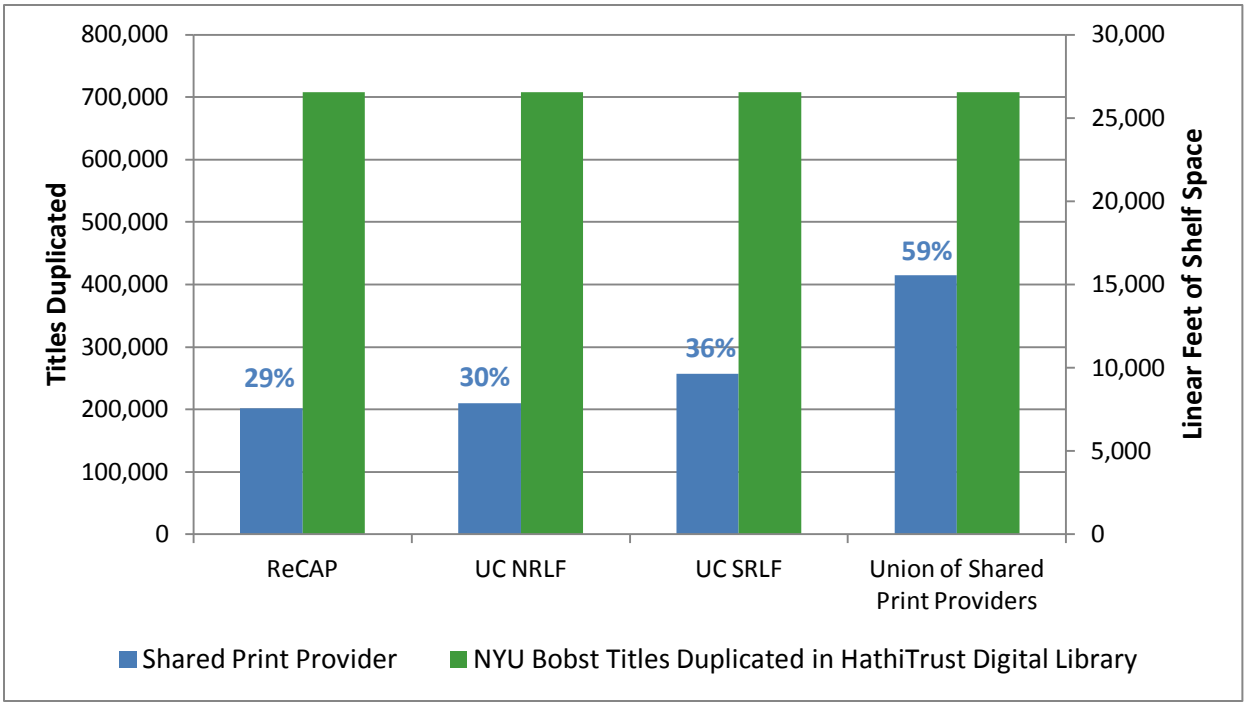


Figure 19. Comparison of potential shared print provision options for NYU Bobst Library (June 2010)

Even if one were prepared to accept the additional challenges of ensuring timely on-demand physical delivery from one side of the continent to the other—which might reasonably be addressed in a shared service business model—it is difficult to imagine that a marginal gain of fifty thousand titles or a few thousand linear feet of shelf space would motivate a library like NYU to go farther afield for a shared print partner.

A potentially more profitable alternative would be to contract with multiple shared print repositories to increase the scope of coverage and maximize the externalization of high-cost, low-return print operations. Yet, as shown in figure 19, comparison of Potential Shared Print Provision Options for NYU Bobst Library, even ***the combination of multiple large-scale***

repositories results in only a relatively modest increase in the tangible value of the partnership: the union of three very large shared print collections scarcely suffices to replace 60% of the mass-digitized titles in NYU's Bobst Library. A recent study (Payne, 2007) identified approximately 70 high-density academic library storage facilities in North America; each one of these repositories may be said to represent a potential shared print service provider. Multilateral agreements could theoretically be struck across this network of repositories to maximize the scope of shared print agreements. At present it is virtually impossible to assess the carrying capacity of this infrastructure, since only a small number of high-density storage collections are currently visible in national and international union catalogs. Until the latent value of this aggregate resource is effectively and systematically disclosed, it will be difficult for individual libraries to judge the benefit of prospective shared print partnerships.

Optimizing Existing Infrastructure

Based on our necessarily limited view of existing infrastructure, there is a significant gap between the level of shared service provision a client library like NYU might reasonably seek and what repositories like ReCAP or RLF might readily provide. This does not reflect an intrinsic flaw in the library system; it is the necessary and natural outcome of a business model in which print collections are acquired and managed primarily as a local resource. In most instances, withdrawal and storage transfer decisions are reactive responses to local space crises and not intentionally guided by long-term library strategy. As a result, off-site depository collections as currently constituted have only limited value as a source of surrogate print preservation and access services for "external" consumers. Their business value as a cooperative resource is correspondingly small.

Ultimately, the benefit and business value that shared print repositories can deliver will be determined not by present inventory or service capacity, but by their individual and collective ambition to transform the academic library enterprise. If the market for shared service is sufficiently great, even commercial interests may be motivated to acquire and manage print inventory on behalf of academic libraries. More probably, some number of existing library repositories will opt to reconfigure collections and operations to support shared service provision, so that academic institutions can outsource management of low use print holdings. This would effectively result in an optimization of the existing library infrastructure as resources are pooled to meet collective service requirements.

We can judge the potential impact of this reorganization by returning to the now familiar NYU use case. If we expand the scope of analysis beyond ReCAP itself (as represented in our sample) to include the totality of holdings in the ReCAP partner libraries—on the presumption that any title held by Columbia, Princeton or NYPL might eventually be transferred to the shared repository—we find that NYU could potentially outsource print management of more than 90% of the mass-digitized titles in its collection. Moreover, comparing figures 16 and 20, one can see that over the twelve months of our study, the ReCAP libraries were consistently capable of supplying more than 90% of the mass-digitized titles in NYU's collection. Simply put, ReCAP has the potential to satisfy NYU's anticipated shared print service need.

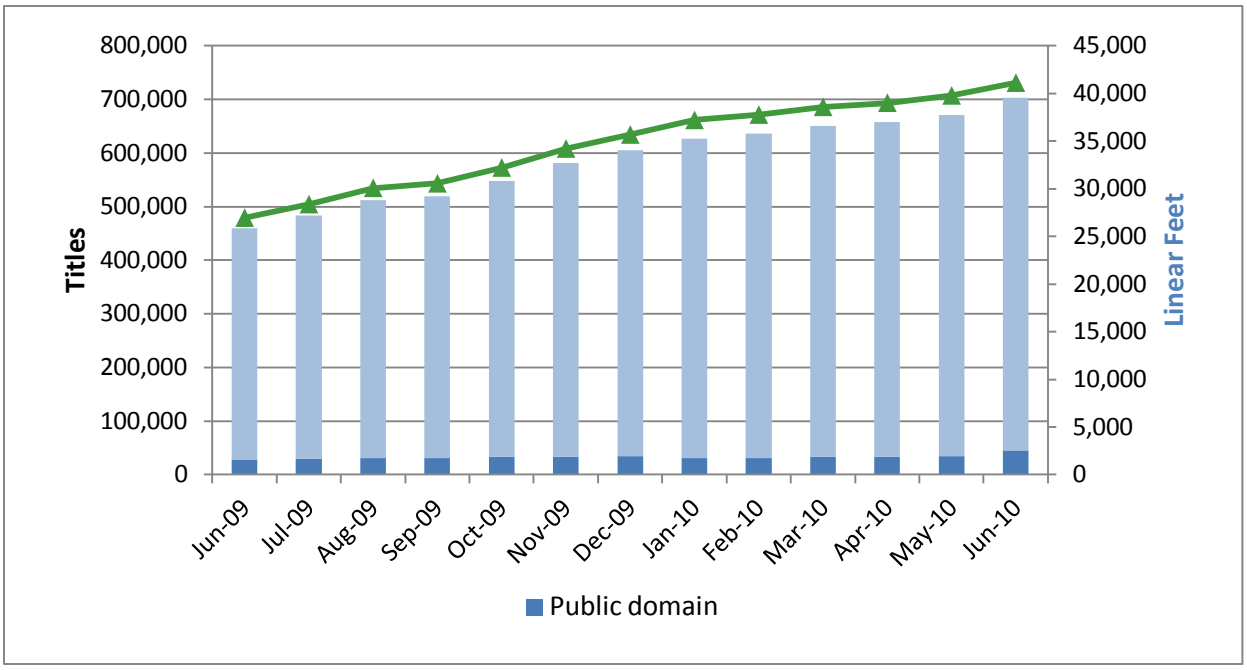


Figure 20. NYU Bobst titles duplicated in ReCAP partner libraries and HathiTrust Digital Library (June 2009-June 2010)

One might conclude from this that NYU would be best served by striking a shared service agreement with the ReCAP partner libraries directly: why wait for the desired inventory to be transferred to the storage facility?

Practically speaking, it would be difficult for NYU to negotiate with the ReCAP consortium for guaranteed access to print collections still managed on-site at Columbia, Princeton and NYPL. The ReCAP consortium was formed to provide a governance structure for the shared library storage facility and does not exercise any joint authority over the partner library collections. NYU would instead need to negotiate individual agreements with each of the partner libraries. Beyond the administrative overhead of negotiating and implementing multiple agreements, there are other factors to consider. First, collections managed on-site have a

higher loss rate than collections managed in a high-density repository and are subject to variable environmental control; the preservation value of individually negotiated agreements will therefore be less than ReCAP can provide. Secondly, the direct cost of managing on-site collections is substantially greater than managing collections in storage, so the shared access agreements would likely be more expensive as well. Finally and most importantly, ReCAP partner collections are not managed as a cooperative resource and as a consequence NYU could not obtain any meaningful assurance that the shared collection(s) would be subject to similar terms and conditions.

As an alternative, NYU could conceivably negotiate a shared print agreement with the ReCAP consortium members that is contingent on the prospective transfer of materials to the shared storage facility. In addition to securing NYU guaranteed access to these materials under common terms and conditions, an agreement like this would deliver benefit to the consortium members by creating an incentive to implement a genuine cooperative management plan that would hasten the transfer of materials to storage (reducing inventory management costs for individual contributors) and potentially reduce or eliminate duplicate deposits (extending the useful life of the facility). Moreover, inasmuch as titles held in common by NYU and ReCAP members are also likely to be held by other academic libraries in the region, accelerated transfer of these materials to ReCAP would increase the likelihood of a scalable shared print business model.

Striking an agreement of this kind, which explicitly articulates expectations of prospective and targeted growth in a shared print service collection, and also leverages the full value of repository infrastructure, would undoubtedly be challenging. The initial gains for NYU and for ReCAP might be relatively modest, but the long-term benefits could be genuinely transformative. NYU, Columbia, Princeton and NYPL could all reasonably expect to regain significant library space and reduce redundant expenditure while contributing to the creation of an optimized and sustainable print collection. Such an agreement might also serve as a model that could be adapted for use with other large-scale storage repositories, expanding the total market for shared print service and helping to establish common norms and expectations. For these reasons alone, a valiant effort is arguably justified.

What is it Worth? Putting a Price on Shared Collection Services

Thus far we have characterized the value of shared collection services primarily in terms of library space recovery: linear feet of shelving that might be freed up to accommodate new acquisitions or assignable square feet that might be repurposed as study space, learning and research commons, etc. However, a deliberate strategy to externalize collection management functions for low-use print inventory can also generate economic benefit in the form of cost avoidance for consumer libraries and cost recovery for shared service providers. Even more

importantly, joint business agreements that effectively redistribute the costs and benefits of print and digital preservation have the capacity to transform the academic library enterprise, freeing individual organizations to pursue service goals that are relevant to local needs while still meeting collective stewardship obligations.

In real terms, we can estimate the economic value that might be obtained through an externalization of print management functions with a simple (and admittedly oversimplified) cost calculation. For a potential consumer such as NYU, the total cost avoidance will be determined by the number of volumes that are covered by a shared print service agreement. Based on the analysis described above, we judge that under present conditions ReCAP could potentially provide surrogate collection services for about 200,000 mass-digitized titles in NYU's Bobst library. Since nearly all of the titles are monographic publications, we will conservatively estimate that each title represents a single volume in the Bobst collection. Using an estimate of $4.26 per volume, we calculate the cost to manage these titles on-site in the Bobst library to be approximately $850,000 per year. In an efficient high-density storage environment like ReCAP, the annual cost of keeping the same number of volumes is significantly less: about $172,000 based an estimate of $0.86 per volume. By outsourcing inventory management functions to ReCAP, NYU might therefore achieve a significant reduction in the per-unit cost of preserving these books, in addition to regaining about 13,000 linear feet of library shelving or 16,000 ASF of space in Bobst.

Of course, since the cost estimate for keeping a book includes a number of "sunk" costs, NYU will not actually recover an amount equal to—or remotely approaching—$850,000 each year. The retail value of the ReCAP service offer is more likely to approximate the savings NYU will realize by not transferring these titles to its own storage facility. Because NYU leases a facility and has thereby avoided the up-front capital costs of construction, the per-volume lifecycle management costs are almost certainly less than $0.86 per year. It stands to reason that NYU would therefore reject a shared print service proposal for which the cost exceeds $170,000 per year. Normal ReCAP operating expenses are already subsidized by the three consortium members, so the marginal costs of offering service to NYU and other libraries would likely serve as the basis of a pricing model, allowing for future growth in the scope of the service collection and projected opportunity for space savings at the client libraries. An ambitious shared print service provider could maximize cost recovery by building a service collection shaped to the needs of an external clientele, substantially reducing or even eliminating the charge backs customarily used to sustain shared storage repositories. For privately funded organizations like ReCAP, as well as publicly funded entities like the UC Regional Library Facilities, the external market for shared print service may represent a path to long-term sustainability.

This is not the place for a comprehensive examination of business models that might support shared print service; logically, that work will be taken up by organizations that aspire to serve as service providers, in consultation with motivated consumers. Instead, we can offer a few tentative observations based on findings from our empirical study of existing infrastructure and anticipated service requirements. If shared service provision is to be developed on the backbone of the existing storage repository infrastructure, as seems likely in the near term, it will be necessary to strike a balance between the need to minimize retrievals from high-density facilities, which would tend toward an "insurance-only" access model, and the interest in maximizing reliance on external providers, which would tend to concentrate demand on a small number of suppliers. Both of these goals could be accommodated in an arrangement in which pricing is determined in part by the demand profile of the service collection.

Returning to the example of NYU, we can anticipate that demand for the 200,000 titles that might be covered in an initial agreement with ReCAP, will vary according to more or less predictable patterns. A pricing scheme that is sensitive to this variability would protect ReCAP from the negative cost consequences of increased retrievals while providing NYU the guaranteed access it requires. Since overall demand is likely to be low across the service collection, a transaction-based pricing model seems inadvisable; ReCAP or any other large library storage repository would find it difficult to generate a reliable stream of cost recovery even if shared access agreements were struck with a large number of client libraries in the region. Instead, an annual baseline contribution from client libraries, pro-rated according to the size and value of the service collection, might offset normal operating expenses, while a variable fee based on the specific demand profile of the titles covered by the agreement would provide necessary flexibility.

Aggregate demand patterns—for example, the higher circulation rates that are typically observed (among North American libraries) for English-language publications of relatively recent vintage, or the extremely low use profile of monographs with non-roman scripts—could be used to establish the overall demand profile of a service collection. This would allow for standardization of demand-based pricing, which would improve the transparency of shared print business agreements. This in turn might stimulate healthy competition amongst shared print service providers. If well-publicized aggregate demand patterns were used to establish pricing rates, individual shared print repositories could establish a competitive advantage by offering comparable service at varying price points.

Discriminatory pricing based on demand profile would allow shared print providers to tailor service agreements according to the needs of institutional subscribers. In the case of NYU, more than half of the titles for which a shared service agreement might presently be struck with ReCAP represent English language monographs published in the last decade. The higher

demand profile for these titles compared other ReCAP holdings would reasonably justify a higher service cost. In theory, pricing and service level agreements might be exquisitely sensitive to variations in demand; practically speaking, a simpler model is likely to prevail, if only because library organizations are not especially entrepreneurial in nature. Ideally, a demand-based pricing model would allow for periodic adjustments based on changes in the library system as a whole, so that the market for shared print service is not subject to dramatic fluctuation, which might have devastating consequences for individual libraries that find themselves priced out of the local marketplace.

From a consumer perspective, the operational value of a shared print agreement offer will be determined by the initial scope and size of the service collection, and its rate of growth over time. Libraries contracting for shared print services will want to achieve maximum benefit in the form of local space recovery and cost avoidance, which is dependent on the scope of the service collection. They will also reasonably seek an assurance of continued growth in the service collection, since a library's ability to derive ongoing benefit from the arrangement requires that new space savings be gained each year. From a service provider perspective, the costs of delivering shared collection service will be determined by the rate at which material destined for a service collection is accessioned and the rate at which it must be supplied. ***It is in the mutual interest of shared service providers and consumers that repository collections rapidly assume the profile of "optimal" service collections, so one can anticipate that prospective providers will as a matter of course begin to accession inventory according to market needs***.

It is worth considering that the increased discoverability of the mass-digitized book collection may result in greater demand for the print version, especially in the absence of a licensing agreement for the in-copyright titles. A recent study of post-digitization use of print collections at the University of Michigan found that the increased discoverability of books made available as full-text resources online did not result in increased demand for locally held print versions (Look, 2010). The scope of the study was small and only addressed titles in the public domain so it is not possible to infer that demand for the much larger in-copyright corpus will be similarly unaffected by increased network visibility. Further analysis of aggregate demand patterns for titles already in storage could provide useful insights into the likely impact of pooling supply and demand for low- and moderate-use academic print collections on a regional basis.

Because retrievals are the single greatest cost driver in high-density facilities, repository managers are motivated to control demand for physical inventory by accessioning only (or mostly) low-use titles. In a shared service context, there is some risk that the concentration of demand from multiple institutions will result in increased retrievals and higher operating costs. This risk could be mitigated by aligning shared print service collections with the mass-

digitized book corpus in Hathi and ensuring that digital surrogates are the primary mode of discovery and delivery. This alignment would serve a dual purpose by maximizing the benefit libraries can derive from the mass-digitization enterprise while also providing a means of moderating physical retrieval rates. The result would be a virtuous circle of shared service provision, in which collective library investment in the creation of the HathiTrust Digital Library is repaid by the increased efficiency in library operations enabled by cooperative print management.

Who Will Benefit? Who Will Pay?

We have established that a deliberate reorganization of the existing ReCAP collection in which inventory is more closely aligned with the growing corpus of mass-digitized texts, along with other dual format titles, would substantially improve its ability to function as a shared print service provider for NYU. We have also examined the distribution of library holdings in both Hathi and ReCAP and hypothesized that there is a substantial market for shared service based on the many hundreds of thousands of titles for which aggregate library holdings are relatively abundant and demand is low, and for which a shared service provision based on existing repository holdings appears feasible. Is the potential market for service sufficiently large to sustain shared print service at scale? Can a core segment be identified for which a common model of service provision might be satisfactory?

To answer these questions, we returned to the constituency from which this project was born: university-based academic research libraries in North America. Measuring the percentage duplication of titles in each of the 113 ARL university libraries and the HathiTrust Digital Library at twelve-month intervals, we established a baseline against which our findings for NYU could be compared. As shown in figure 21, the results indicate remarkably low variance in duplication levels across the ARL cohort. This is an especially notable finding since there are great disparities in the library volume (and respective title) counts among ARL libraries, ranging from more than 16 million volumes at Harvard University to fewer than 2 million volumes at the University of Guelph, based on data reported to ARL in 2007-2008. In June 2009, an average of 20% of titles held in any given ARL library was duplicated in the HathiTrust Digital Library; by June 2010, the average duplication rate had increased to 30%. These figures are consistent with the levels we found for NYU's Bobst Library.

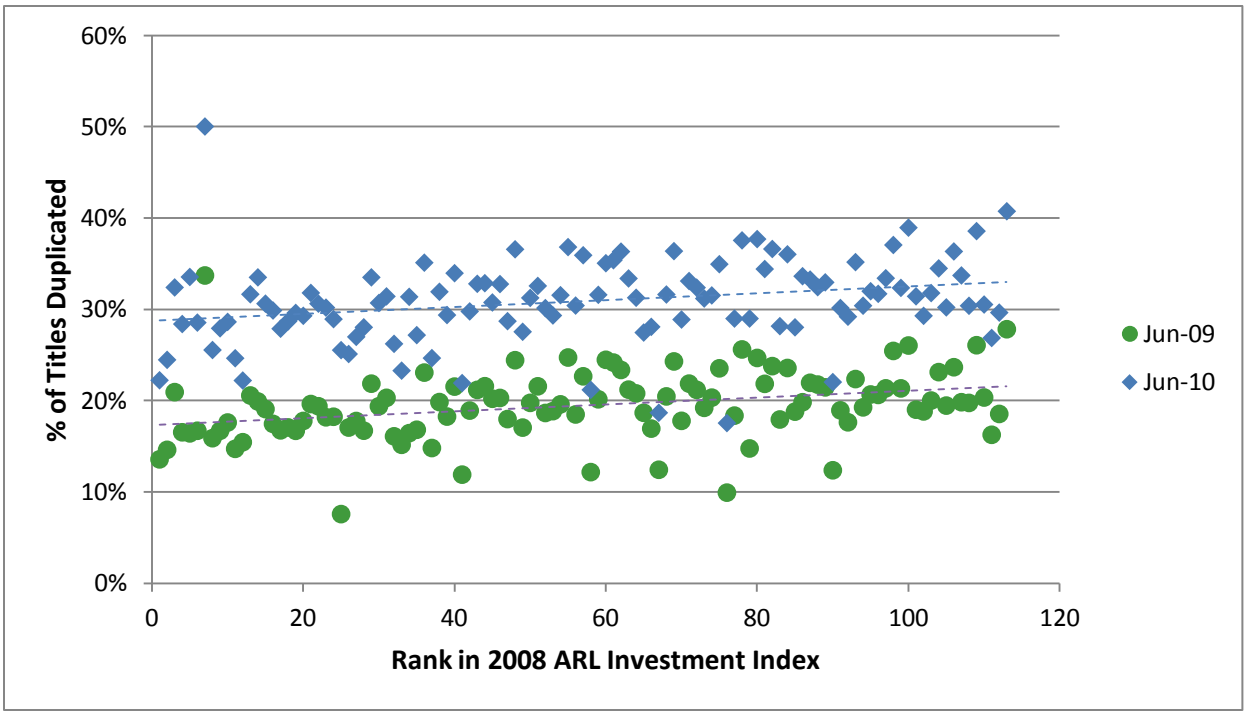


Figure 21. Percentage duplication of titles held in ARL libraries and HathiTrust Digital Library (June 2009 and June 2010)

This scatter chart provides a simple but effective visualization of an important pattern that this project has revealed: that is, that the risks and opportunities associated with moving collection management "into the cloud" are uniformly distributed across the research library community as a whole.

Based on these findings, we estimate that the median space savings that could be achieved at an ARL library if a robust shared print offer were in place today to be approximately ***36,000 linear feet or the equivalent of more than 45,000 ASF***. To put this in perspective it is helpful to consider that the space requirement for a typical library-based research or learning commons is about 20,000 square feet. In economic terms, the total annual cost avoidance—assuming all of these books are currently managed on-site—***exceeds $2 million per library***. As noted above, this is not a sound basis for judging how much library resource would actually be available for redirection in support of other operations; however, it does provide a useful measure of the opportunity costs of inaction. The cost of managing these same titles in a high-density off-site facility would amount to approximately $500,000 per year for each library. This provides a rough measure of the maximum retail value for a shared print offer that would enable a full externalization of print operations for mass-digitized books.

These figures represent a necessary oversimplification of what is obviously a much more complex and challenging business case. For example, a significant number of ARL libraries own off-site shelving facilities and it is likely that some of the mass-digitized titles that might otherwise be relegated in favor of a shared service agreement are already "locked up" in storage. The calculation of benefit for those libraries would necessarily be different. By the same token, some libraries that have already transferred a significant portion of the locally owned inventory to storage may see themselves as potential shared print providers, rather than consumers. It is not possible to provide an accurate forecast of which libraries will likely assume a role as a supplier or consumer; nor it useful to speculate about those that will disdain the shared service model, preferring instead to operate in relative isolation. Still, we feel that our broad brush estimate of the market for shared print service in the ARL community provides a useful starting point for library directors and repository managers to begin planning and even budgeting for a future in which managing local print collections is no longer a core function or cost center.

Conclusions and Recommendations

Our year-long study of the mass-digitized book corpus in the HathiTrust Digital Library and parallel investigation of potential shared print service providers has confirmed that there is an opportunity for significant library space savings and cost avoidance if management operations for digitized books are deliberately and systematically outsourced or externalized to shared service providers.

One can anticipate that academic institutions interested in reducing local print holdings in favor of regionally consolidated inventory will, in years to come, increasingly look to extant repositories like ReCAP, the UC Regional Library Facilities, etc., as a source of preservation and access services. Our findings suggest that current storage inventory is not presently optimized to support shared print solutions on a large scale, but also indicates that system-wide reorganization of collections and services that maximizes the business value of print as a cooperative resource is both feasible and capable of producing great benefit to the academic library community.

Our findings suggest that the shared infrastructure that is needed to support a broad-based externalization of legacy print management functions is unlikely to materialize without some purposeful action by the academic library community. By describing and—where possible—quantifying the value that a changed infrastructure might deliver, we hope to have contributed in some measure to stimulating potential consumers and suppliers alike. Further work will be needed before academic libraries and the educational institutions they serve can fully realize the benefits of shared service:

- It is in the interest of all academic libraries that mass-digitized collections be made more widely available to students and researchers, and that their scope and quality improve to the degree that low-use print inventory can be retired in favor of increased reliance on digital surrogates. Library directors and academic administrators should advocate in favor of licensed access to the mass-digitized resource as part of a comprehensive strategic plan in which the library can reassert its role as a vital part of the academic enterprise.

- The HathiTrust's ongoing efforts to expand public access to the mass-digitized book corpus through programmatic rights assessment, direct negotiation with rights holders, and by accessioning large aggregations of digitized public domain resources, should be recognized as a major contribution to the transformation of the library service environment. By developing a partnership model that is not dependent on content contribution, Hathi can deliver benefit to an even broader range of academic institutions.

    - *Beginning in 2013, the HathiTrust will introduce a new cost model that will enable non-content-contributing members to participate in (and benefit from) joint stewardship of the digital repository. This new affiliation model is described in appendix I.*

- Institutions and organizations that aspire to roles as shared print service providers need to proactively build collections that will deliver maximum operational value to external audiences; they should leverage the collective library investment in Hathi by accelerating the transfer of mass-digitized titles to print preservation repositories and self-consciously promote these resources as shared and cooperatively managed assets. Above all, these institutions must take steps to make their service ambitions and capacity known, so that potential consumers can begin to articulate core requirements and common service profiles can be identified.

    - *In cooperation with other library organizations, OCLC is working to develop technical solutions that will enable the latent value of library storage collections and distributed print archives to be more effectively disclosed.*

- Prospective shared print providers can help to define a common service profile by surfacing model agreements and engaging in dialog about the operational and business requirements of shared service provision. Managers of large print repository collections should be empowered and encouraged to engage in business modeling exercises that are explicitly intended to expand the market for service beyond content contributors.

  - *ReCAP partner libraries have outlined core elements of a shared service agreement. These are summarized in appendix II.*

- Libraries that are already motivated to outsource legacy print management functions in support of a changed service portfolio or simply to relieve local space pressures should begin to establish objective targets, and quantify and articulate desired outcomes so that motivated suppliers can respond in kind. Library administrators should engage directly with faculty and academic officers to communicate a compelling strategy in which selective externalization of traditional functions is demonstrably improving the institution's ability to fulfill an academic and research mission. This work will be challenging and deserves external support and endorsement by library leadership organizations and funders.

- Research organizations can advance our collective understanding of the changing profile of demand for legacy print in the mass-digitized environment and help to characterize the optimal redistribution of library resources.

When these steps are taken, we will have made measurable progress toward the worthy goal of ensuring the long-term survivability of the scholarly record at a cost that is sustainable for the research library community as a whole.

Appendix I. HathiTrust Cost Rationale

Beginning in 2013, HathiTrust will use a cost model that reflects benefits that partners can receive from works stored in HathiTrust rather than the cost associated with storing them. We believe that this new cost model, focusing on consumer values rather than storage costs, better reflects the long-term interests of partners and will more fairly distribute costs across the partners.

This new benefits-based cost model attributes the cost of storing a volume to each library that holds (or held)[3] a corresponding print volume. In this model, a library that pays for a share of the cost of storing content is also acknowledged as receiving the benefit of the content. Those benefits for in-copyright volumes are not always tangible, but include producing replacement copies and offering specialized services permitted by contract or law. Such a cost model would, for in-copyright works, attribute a share of costs to all partnering libraries that hold or held the corresponding print volume. Because all member libraries enjoy the benefits of public domain works, every partnering library will be assumed to hold these works.

This new model of cost attribution will, compared to the current storage-based model, have a smoothing effect, reducing the cost borne by an institution that contributes significant content, and recovering cost from each member of a new class of partner libraries ("Sustaining Partner" libraries) that shares in the benefit of these volumes.[4]

The current HathiTrust membership is comprised primarily of institutions contributing large amounts of content and, thus, bearing large costs. At this time (late 2009), we are in conversation with research libraries that do not have large amounts of content to contribute, but wish to join HathiTrust to participate in its curatorial work. Under the current model, partners with smaller amounts of content pay relatively small amounts and have access to large amounts of content. By applying a model based on shared holdings, we will see some reduction in the cost per contributing institution as more libraries join the effort. Clearly, as these new libraries join, it must be with the understanding that a new cost model will work to distribute these costs in an equitable way that reflects the benefits accruing to all partners.

This holdings-based cost model will incorporate a number of precise elements about costs and the "sharedness" of the content. The cost will be recalibrated each year, as costs for infrastructure will change each year. It is also the case that the current partners would like to be able to use shared funds to develop new services and functionality. Consequently, we will multiply the cost of infrastructure by a variable amount (adjusted periodically) to fund those new services and functionality. In this new holdings-based cost model, the costs to an institution per year will be calculated as follows:

For public domain volumes:

(PD*X*C)/N

where

- PD is the total number of public domain volumes in HathiTrust (assumed to be "held" by all partner libraries). This number will also include in-copyright works where the rights holder has given the members free use of the content.
- C is the average annual cost to provide basic support for a volume. Note that costs will vary by volume, as each printed volume will vary in number of pages and in average file size. We will use an average cost that will be periodically recalculated.
- X (greater than one) is a value with which we multiply C to generate a surplus.
- N is the total number of partner libraries.

For a given in copyright volume, IC:

IC = (C*X)/H

where

- C is the average annual cost to provide basic support for a volume (as above).
- X (greater than one) is a value with which we multiply C to generate a surplus.
- H is the number of partner libraries that hold a given print IC volume.

Initially, HathiTrust proposes using a value of two (2) for X, i.e., doubling the cost of maintenance in order to build a fund for services. Thus, 50% of the funds collected will go to development and the other 50% will cover the costs of storing content, again lowering cost proportionally for institutions that contribute large numbers of commonly held volumes.

We believe that this new model will be both equitable and sustainable, but acknowledge that it also presents significant challenges. The biggest of these is the lack of information in the library community currently, on a volume-by-volume and institution-by-institution basis, about overlap in our print collections. No organization, including OCLC, stores information about holdings, and even where OCLC succeeds in approximating this, it lacks volume-specific information. We also face challenges with regard to reliable data. For example, although each of our institutions individually stores volume-specific information, enumeration and chronology information is represented so variously that manual remediation will be required to make it uniform. Ensuring that this information is up to date is an additional concern. HathiTrust, either by itself or in collaboration with another organization, will attempt to create a system to store partner holdings information in such a way that it can be updated constantly, by partner institutions themselves or by central HathiTrust staff.

We believe that this volume-specific infrastructure will be valuable for a number of purposes, including:

- **De-duplication**: although duplication of contents is not costly, storing duplicates compromises the user experience and it obscures collection development needs;
- **Management of corresponding print volumes**: how will we know that we can withdraw a print journal without having volume-specific information?
- **Legal uses of in-copyright materials**: for example, Section 108 uses will depend on having a clear sense of which institutions own(ed) which print volumes.

A clear sense of volume-specific information for digital materials and corresponding print volumes will be needed as our collaboration in HathiTrust develops. HathiTrust plans to launch this new cost model in 2013, during the second phase of our initiative (i.e., subsequent to the first five years of HathiTrust). Prior to that time, we will work to develop the necessary infrastructure to be able to perform these calculations reliably. We will also, effective immediately, entertain membership from other research libraries that wish to share this curatorial role. To calculate costs for these Sustaining Partner libraries, we will use overlap formulas developed in our current explorations of this model with RLG, ReCAP and New York University Library. All funds generated through the participation of Sustaining Partner libraries prior to 2013 will be devoted the development of the new common holdings infrastructure.

jpw, 12 Feb 2010

Appendix II. Cloud Library Service Agreements: ReCAP as Shared Print Repository

DRAFT FOR DISCUSSION

Assumptions:

The objective is to define a service agreement under which ReCAP will provide access to a defined set of print materials under specified conditions, allowing other libraries ("clients") to discard their own print copies. Several basic assumptions inform this model, but each leaves room for some variance in interpretation.

1. A digital copy will be readily available to the client library's patrons.
2. The agreement applies only to copies once owned by the client library and since withdrawn, i.e., the agreement is intended to allow client libraries to save the expense of storing print materials, not to expand the client's collections, or to serve as a back-up to its print collection.
3. The agreement must provide reasonable assurance of continued access to the defined materials. It will therefore limit the freedom of action of ReCAP partners to remove materials from ReCAP permanently or place future restrictions on use.

4. The agreement will not restrict the use of the defined materials by the owning ReCAP partner's own patrons, nor will it carry an obligation to replace items lost or damaged through normal use, i.e., the level of assurance of long-term access will be comparable to that which would be provided for the client's own unrestricted collections.

5. The agreement must provide a level of access and service greater than that available through standard interlibrary loan.

From the client perspective, there is a further assumption: that the digital copy will be preserved and will continue to be accessible. In the Cloud Library project, the implications of this assumption are being tested through a model service agreement with the Hathi Trust. From the ReCAP perspective, the key issue is continued availability of the digital copy to the client library's patrons. The means of assuring that availability is relevant only to the extent that it defines the scope of the agreement (discussed further below).

Elements of an agreement and related issues:

General Considerations:

Any agreement might be closely defined and tightly construed, or might be left relatively open, allowing for greater flexibility. Similarly, an agreement might place greater obligations on ReCAP or on the client, not only with regard to the service provisions, but also with regard to the steps needed for implementation. As an experiment with a new model, an agreement between ReCAP and a single client might benefit from greater flexibility, both in definition and execution. Ultimately, ReCAP would want to have the same agreement with multiple clients, and a single library might want agreements with several repositories. That would argue for closer specification of terms.

Governance:

In the Cloud Library project, ReCAP serves as a representative regional repository. In reality, the form of a service agreement would be shaped to some extent by the governance structure of the repository. At present, an agreement with "ReCAP" would place obligations on both ReCAP staff and ReCAP partner libraries and would thus need approval by the ReCAP Board. Ideally, the agreement would cover the collections of all ReCAP partners on similar terms, but there would be nothing to preclude an agreement that applied only to one or two partners' materials. Once a model agreement had been implemented, the ReCAP Board might decide to authorize the ReCAP Director to execute similar agreements with other clients.

Any agreement intended to cover a partner's collections outside of ReCAP would need separate agreements with the partner library (governing scope and policies) and with ReCAP (governing services to be provide by ReCAP itself).

Scope:

As originally envisioned, the Cloud Library project was limited to books held at ReCAP and by NYU, and accessible through Hathi Trust (i.e., books in public domain). At present, these stringent requirements apply to only a small percentage of NYU's collections. The value of an agreement could be extended if it were expanded to cover:

- All Hathi Trust books held at ReCAP and by NYU, regardless of copyright status, if the digital copy is accessible by other means (such as a Google Book Search subscription database).
- All of the ReCAP Partners' collections also held by NYU and Hathi, regardless of whether the books are currently stored at ReCAP.
- All digitized books held at ReCAP and by NYU and available through any trusted digital repository (e.g., Portico).
- Any combination of the above.

Each of these expansions would affect the nature and terms of any service agreements.

Terms of Use:

As noted above, the client library would want level of access and service greater than that available through standard interlibrary loan. This may be achieved through a combination of several factors:

a. The ability for the client's patrons to discover holdings and place requests via the client's catalog;

b. Expedited delivery;

c. Extended loan periods;

d. Availability of materials excluded from general interlibrary loan.

Ideally, the client would want a level of access similar to what it would provide through a locally-managed remote storage facility.

ReCAP partner libraries would want to ensure that extending access to client libraries would not cause a significant deterioration of service to their own patrons. As an experiment, ReCAP might be willing to extend liberal terms of use, on the assumption that ready availability of digital copies will further reduce demand for these already low-use books, both by the clients' patrons and by the owning library. An initial agreement might include provisions for monitoring the level of use and adjusting terms if necessary.

Operational issues and responsibilities:

Defining Eligible Materials:

The agreement would broadly define the scope of materials covered. It might, for example, cover:

a. all books stored at ReCAP as of the date of the agreement or thereafter, except those with borrowing restrictions, if;

b. the book was also owned by the client as of the date of the agreement, and;

c. a digital copy is freely available to the client's users.

Acting on this definition would require the parties to compile, share, and maintain data. Responsibility might be placed with either party, or contracted by mutual agreement to a third party. For a generalized agreement—one applicable to multiple clients—ReCAP might agree to provide a periodic list of eligible books, contracting with OCLC to analyze overlap with Hathi Trust for example. Client libraries might take responsibility for analyzing overlap with their own collections, or might contract that to ReCAP (and indirectly to OCLC) for an additional fee.

From the client standpoint, only those books owned as of the agreement date would be considered eligible. So, ReCAP might prefer to maintain a file of those titles, and notify clients periodically of any books that become subject to the agreement as a result of additions to ReCAP or Hathi Trust holdings.

Requesting materials:

ReCAP would be responsible for providing clients with sufficient information to place requests, but "sufficient information" could have different meanings, with different costs:

- ReCAP could supply both bibliographic information and barcode numbers, and require clients to submit requests in a standard format including the barcode number, for automated processing.

- Or, ReCAP could supply only bibliographic information and receive and process requests in a manner similar to that for interlibrary loan.

The latter method allows the client greater freedom, but would incur greater cost. (It should be noted that ReCAP itself is not in a position to supply bibliographic information directly; that information would be compiled from the ReCAP partners or possibly OCLC.)

ReCAP might also assume responsibility for notifying clients when items are in use elsewhere and therefore unavailable. Alternatively, ReCAP could provide this information only when such an item is requested. Given the expected low use, the latter may be more cost-effective.

Delivering Materials:

ReCAP would commit to a specific turnaround for filling requests—most probably, one business day. ReCAP does not currently operate a courier service; instead, each partner is responsible for arranging to pick up and return materials. Interlibrary loans are shipped by UPS. For the Cloud Library agreement, ReCAP might offer several delivery options at different costs. Alternatively, a client might contract separately with one of the ReCAP partners for delivery.

Terms of Use:

A Cloud Library agreement would define the terms under which requested items could be used: length of loan, renewals, right to recall items, etc. Given the expected low demand, ReCAP's partners might agree to terms similar to those extended to their own patrons. More than one level might be defined, allowing some items to be used only on site in the client library, for example, so that the agreement could be extended to items not generally available for circulation.

ReCAP itself does not have any mechanism to control terms of use once an item leaves the facility; in effect, all items are supplied on indefinite loan. It might be left to each partner to devise its own means for enforcing limited loan periods, recalling needed items, etc. Alternatively, these activities might be added to the responsibilities of ReCAP's interlibrary loan staff, at additional cost.

Notes

1. This estimate is based on median figures for Volumes Added (Gross) as a percentage of Total Volumes in Library as reported in the ARL Annual Statistics Tables for the five years from 2003/2004 through 2007/2008. http://www.arl.org/stats/annualsurveys/arlstats/statxls.shtml.
2. A mapping of OCLC Conspectus divisions to respective Dewey Decimal, Library of Congress and NLM call numbers is available here: http://www.oclc.org/collectionanalysis/support/conspectus.xls.
3. Particularly for Section 108 uses, a library may wish to withdraw (and thus no longer hold) a volume. We should store information that shows that this library once held the volume in question.
4. The philosophy of collective costs and collective holdings already underpins much of the CIC approach to HathiTrust.

References

Connaway, Lynn Silipigni, Edward T O'Neill, and Chandra Prabha. 2007. "Last copies: What's at risk?" *College and Research Libraries*, 68 (4): 370.

Courant, Paul N., and Matthew "Buzzy" Nielson. 2010. "On the Cost of Keeping a Book." In: *The idea of order: Transforming research collections for 21st century scholarship*. Washington, D.C.: Council on Library and Information Resources. http://www.clir.org/pubs/reports/pub147/pub147.pdf.

Lavoie Brian, and Lorcan Dempsey. 2009. Beyond 1923: "Characteristics of potentially in-copyright print books in library collections." *D-Lib Magazine,* 15 (11-12). http://www.dlib.org/dlib/november09/lavoie/11lavoie.html.

Look, Helen. 2010. *Mass digitization: analyzing online vs. print usage at a large academic research library*. http://www.arl.org/bm~doc/LookPoster.pdf.

Michalko, James, Constance Malpas, and Arnold Arcolio. 2010. *Research libraries, risk and systemic change*. Dublin, Ohio: OCLC Research. http://www.oclc.org/research/publications/library/2010/2010-03.pdf.

Payne, Lizanne. 2007. *Library storage facilities and the future of print collections in North America*. Dublin, Ohio: OCLC Programs and Research. http://www.oclc.org/programs/publications/reports/2007-01.pdf.

Schonfeld, Roger C., and Ross Housewright. 2009. *What to Withdraw? Print Collections Management in the Wake of Digitization*. [United States]: Ithaka S + R. http://www.ithaka.org/ithaka-s-r/research/what-to-withdraw/.

An Art Resource in New York:
The Collective Collection of the NYARC Art Museum Libraries

Brian Lavoie and Günter Waibel

Introduction

New York is a city rich in art resources, and home to some of the world's great art museums and collections. Visitors to Manhattan can admire paintings by Van Gogh, Picasso, and Matisse at the Museum of Modern Art; inspect the Metropolitan Museum of Art's world-renowned collection of musical instruments; and walk through the galleries of the "old masters" collection at the Frick. A short subway ride to the Brooklyn Museum reveals still more treasures, including a unique collection of ancient Egyptian art.

The paintings, sculptures, and other works of art held in the permanent collections of these and other art museums in the New York City area represent a world-class art resource. The extent of this art resource, however, goes well beyond the artifacts themselves. The four institutions mentioned above collectively hold more than a million items in their affiliated libraries: a collection of books, periodicals, catalogs, and other materials spanning the history of art from the ancient to the modern. While these materials originally were collected for the curatorial staff at each museum, this resource now attracts and supports an international art community of researchers, students, art professionals, and increasingly, the general public. The breadth and depth of this resource is amplified by ignoring the boundaries between individual collections, and focusing instead on the aggregate: in other words, the concentration of art-related information resources available in the New York City area.

Libraries are finding more and more opportunities to extend their perspective beyond the boundaries of the local collection. Studies of aggregate collections—the combined holdings of multiple institutions—have been applied to a range of topics, from thinking about ways to

This paper was originally published in May 2008 as an OCLC Research report.
http://www.oclc.org/research/publications/library/2008/2008-02.pdf.

expand the array of resources accessible to users, to identifying opportunities to improve efficiency and eliminate redundancy. Aggregate collection analysis can confirm widely-held, yet unproven "received wisdom" about the size and characteristics of the collective holdings of a group of institutions, as well as reveal aspects that were previously unknown.

The New York Art Resources Consortium (NYARC) includes the Frick Art Reference Library, the Metropolitan Museum of Art's Thomas J. Watson Library, and the libraries of the Brooklyn Museum and the Museum of Modern Art. NYARC was formed under the auspices of a Mellon Foundation planning grant aimed at exploring opportunities for deeper collaboration among the four libraries. As part of this effort, three of the NYARC members—the Frick Art Reference Library and the libraries of the Brooklyn Museum and the Museum of Modern Art—recently announced the selection of Innovative's Millennium ILS platform to host a new shared catalog offering integrated access to the collective holdings of the three libraries.[1]

This paper reports the results of a study examining the size and characteristics of the aggregate collection of the NYARC member institutions. The goal was to provide these institutions with an empirical context for their ongoing discussion on future opportunities for collaboration. The study also represents a general demonstration of the value and potential applications of aggregate collection analysis. The remainder of this paper is as follows:

- Section I provides a few remarks about the NYARC art libraries, and the data used for this study
- Section II discusses the size and holdings patterns of the NYARC collective collection
- Section III discusses some of the characteristics of this collection, with an emphasis on two material types of special interest to art libraries: exhibition catalogs and auction catalogs
- Section IV examines the degree of overlap of the NYARC aggregate collection compared to the library system as a whole, several other New York-area institutions, and a peer institution located in another part of the country
- Section V draws on conversations with representatives of the NYARC libraries to sketch out some possible applications for this kind of analysis in terms of future planning and decision-making
- Section VI offers some concluding thoughts.

I. A note about the NYARC libraries and data

The motivation for collaboration among the NYARC institutions emerges from both the similarity and distinctiveness across their collections. In terms of similarity, the NYARC institutions share a mission to support their curatorial staff as well as researchers, students, and the general public. Hence, collaboration helps reinforce a shared mission. Despite the differences in the art work collected at each institution, the art museum libraries overlap to some degree in the bibliographic materials acquired, either in regard to classes of materials (e.g., all four institutions collect exhibition catalogs), or even in regard to specific titles (for example, general art reference works and databases). Collaboration among the NYARC institutions therefore helps identify opportunities to remove unneeded redundancy. The benefits from collaboration are also enhanced by differences across each of the four NYARC library collections. Because each museum specializes in different forms of art work, the nature of the art museum library collections will also be different. Collaboration among the NYARC libraries therefore helps them leverage the distinctive features of each library collection across a wider audience.

All of these motivations for collaboration can be encouraged and made more concrete by aggregate collection analysis. The fact that there are similarities across the collecting activities of the four libraries suggests opportunities to minimize redundancy; analysis of the NYARC institutions' aggregate holdings will characterize the degree to which the four collections overlap, and more specifically, help identify areas where redundancy can be usefully eliminated. Similarly, each individual NYARC library collection has a distinctive contribution to make to the combined NYARC resource. Aggregate collection analysis can marshal tangible evidence to support the assertion that the collective holdings of the four institutions embody a resource of greater depth and scope than any single collection in isolation.

The analysis of the NYARC aggregate collection is based on data from the RLG Union Catalog and the SCIPIO database of auction catalogs (prior to the integration of these databases with WorldCat, the OCLC bibliographic database). The data used for the study was extracted in January 2007.[2] The study explores various aspects of the size, scope, and characteristics of the four libraries' "collective collection." While a comprehensive set of results has been shared with the NYARC libraries, this paper focuses on outcomes of general interest. Results attributable to a particular institution are omitted.

II. Size and holdings patterns

The individual collections of the four NYARC institutions exhibit significant dispersion in size: the largest collection is about three-and-a-half times the size of the smallest. While even the

largest NYARC collection is small in comparison to that of a typical academic library, adding the four collections together (without eliminating duplicate holdings across institutions) yields a combined resource of over 1.1 million items. Taking into account that the subject range represented in an art museum library collection is necessarily limited, focusing primarily on materials related to the institution's object collections and the world of art generally, a more appropriate comparison would be to other institutions' holdings in similar subject areas; by this yardstick, the NYARC institutions, individually and collectively, manage a research and learning resource of considerable proportions.

When analyzing the aggregate holdings of multiple institutions, it is useful to eliminate duplicate holdings across institutions in order to achieve a more accurate perspective on how the scope and depth of the institutions' collective holdings expand through aggregation. In light of this, we define the NYARC aggregate collection as the combined holdings of the four institutions, adjusted to eliminate duplicate holdings—in other words, the collection of unique titles held by the four institutions.

With this in mind, the NYARC aggregate collection, as represented in the RLG Union Catalog (RUC) and SCIPIO databases in January 2007, consists of 962,290 unique titles. Eliminating duplicate holdings therefore reduces the size of the four institutions' combined holdings by 17 percent. This suggests that when the holdings of the four institutions are combined, less than one item in five is held by at least two NYARC institutions; the overlap across the institutions is relatively small. As a point of comparison, a recent study of the original five libraries participating in the Google Book Search digitization program[3] determined that combining the print book holdings of the five libraries resulted in a redundancy rate of about 40 percent. We can gain a better perspective on the relative uniqueness of the NYARC collections, both individually and in the aggregate, by taking a closer look at the holdings patterns embedded within them.

More than 80 percent of the titles in the NYARC aggregate collection are held by a single institution, compared to less than 1 percent (or 4,170 titles) held by all four (figure 1). This suggests that a high degree of uniqueness exists across the four individual NYARC collections, which in turn suggests a value in aggregation: the collective holdings of the four institutions represent a collection of far greater scope than any single collection in isolation.

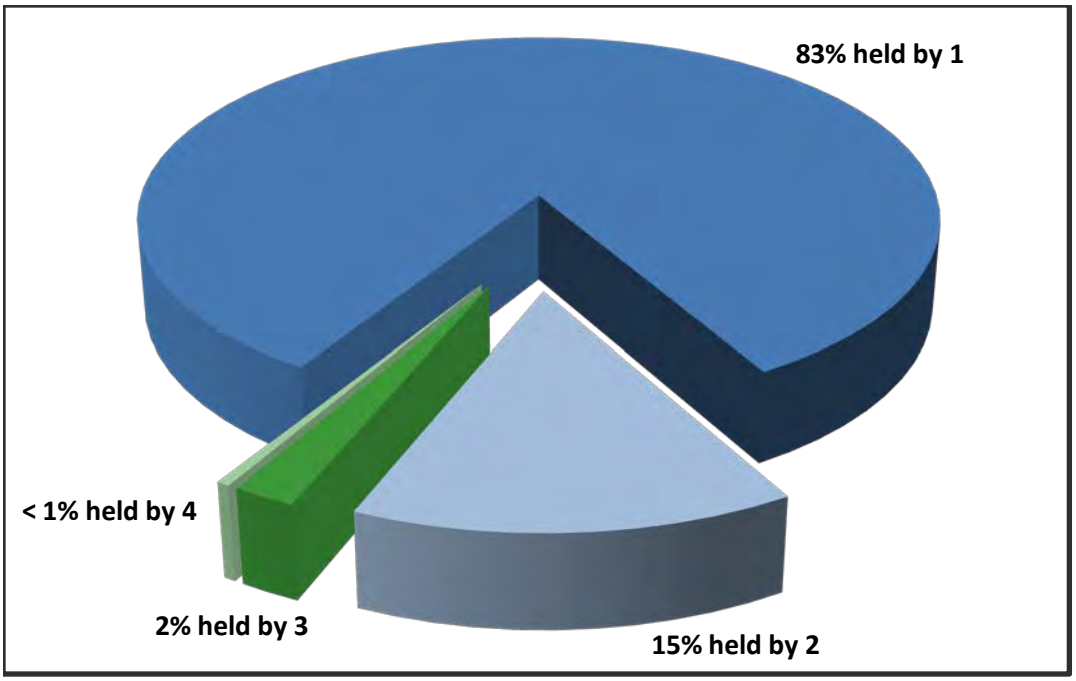


Figure 1. Holdings patterns in the NYARC aggregate collection

Table 1 reports separate holdings patterns across the four NYARC institutions for titles appearing in the RUC database, and those appearing in the SCIPIO database of auction catalogs. Breaking the NYARC aggregate collection down into its RUC and SCIPIO components permits the analysis to isolate holdings patterns for auction catalogs, which are a class of materials of special interest to art museum libraries. Comparison of the RUC and SCIPIO holdings patterns indicates that the latter exhibits a slightly higher degree of overlap than the former, suggesting more convergence in collecting activities in regard to auction catalogs vis-à-vis the remaining materials in the NYARC aggregate collection. However, the difference is small, and the degree of uniqueness of auction catalog holdings is still high, with nearly 80 percent of the auction catalog titles in the aggregate collection held by a single NYARC institution.

Table 1. Holdings patterns, RUC and SCIPIO

Holdings Pattern	RUC Only	SCIPIO Only	Total
Held by one institution	83%	79%	83%
Held by two institutions	14%	21%	15%
Held by three institutions	2%	< 1%	2%
Held by four institutions	1%	N/A*	< 1%

**Note: Only three institutions report auction catalog holdings in SCIPIO.*

In considering uniquely held materials—that is, materials held by a single institution—it is useful to account for materials that are "intrinsically unique": for example, archival materials, or "vertical files" of clippings and other materials organized by artist or gallery. Resources of this kind are by their very nature uniquely held; no other institution could have a precisely equivalent resource. Given this, the question arises as to whether the degree of uniqueness evident across the individual NYARC collections is mainly attributable to these "intrinsically unique" materials. Consultation with NYARC participants in the study yielded reasonable bibliographic criteria for isolating these special materials; subsequent analysis indicated that 117,488 titles in the NYARC aggregate collection fell into this category, or about 12 percent. Holdings patterns for the NYARC aggregate collection, excluding these "intrinsically unique" materials, are shown in figure 2.

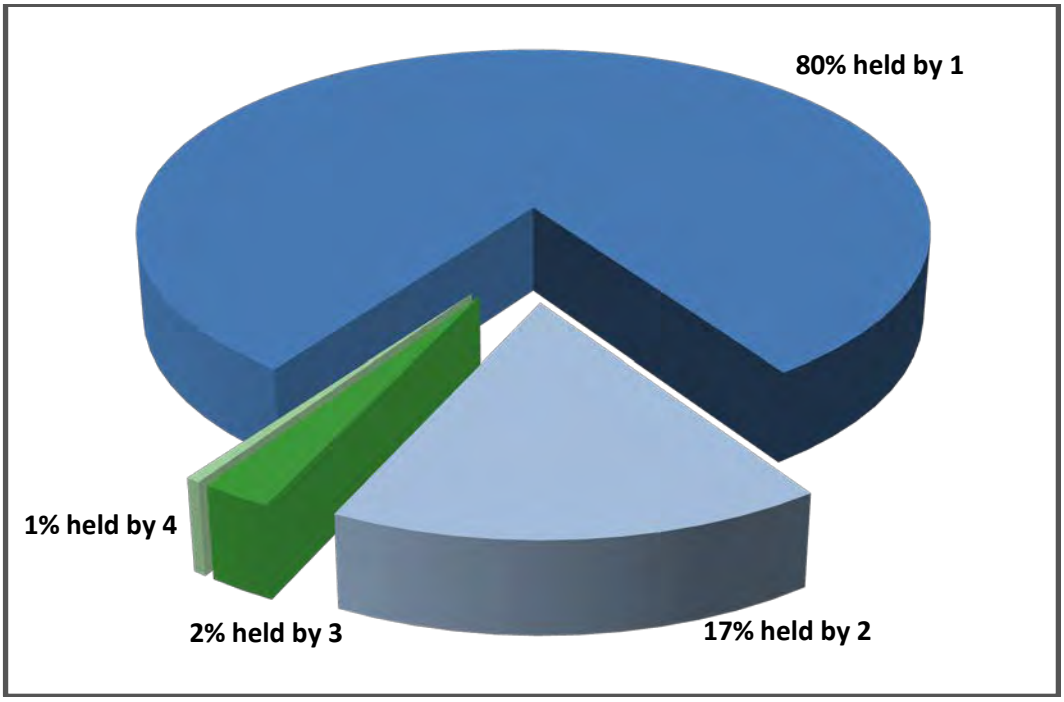


Figure 2. Holdings patterns in the NYARC aggregate collection, excluding "collections"

Excluding these special materials has little impact on the overall contour of the holdings patterns across the four NYARC institutions: 80 percent of the remaining titles are still held uniquely by a single institution. This reinforces the notion that the degree of uniqueness across the four collections is high, and suggests that this uniqueness arises not just from the presence of archival or other special materials in the collections, but also as a product of differences in collecting activity within the realm of published materials.

The high degree of uniqueness present in the combined holdings of the four NYARC institutions suggests another question: is this uniqueness disproportionately attributable to the holdings of one or two institutions, or are the unique contributions to the aggregate collection spread relatively evenly over all four institutions? To answer this question, we computed the percentage of each NYARC institution's collection that was unique relative to the NYARC aggregate collection as a whole; computing the percentage, rather than number of titles, controls for differences in collection size. Results indicated that a significant portion of each collection was unique relative to the aggregate collection, ranging from a high of 79 percent for one institution to a low of 58 percent for another. Excluding the "intrinsically unique" materials from the analysis did not impact these numbers significantly: the proportions decline to a high of 70 percent for one institution to a low of 57 percent for another, but are still highly significant. In short, each of the individual NYARC collections appears to be highly unique compared to the combined holdings of the other NYARC institutions.

Another perspective on the degree of uniqueness within the NYARC aggregate collection is obtained through an examination of pair-wise holdings overlap between the NYARC institutions. In other words, given any two NYARC institutions, what is the degree of overlap between their two collections? Analysis indicated results ranging from a high of 12 percent overlap, to a low of 3 percent overlap. Put another way, no two NYARC institutions exhibited more than a 12 percent overlap across their two collections. As before, removing the "intrinsically unique" materials from the analysis did not impact the results significantly: the range of results adjusts slightly to a high and low of 13 percent and 4 percent, respectively.

These results corroborate the existence of a high degree of uniqueness across the four NYARC collections. The considerable cross-collection uniqueness results in an aggregate collection of far greater scope, depth and utility than any single NYARC collection in isolation.

III. Some characteristics of the aggregate collection

The analysis to this point has focused on the size and holdings patterns of the NYARC aggregate collection. In this section, we analyze the characteristics of the materials in the NYARC collection along a variety of dimensions, including material type, language, and publication date.

Material types

The vast majority of the materials in the NYARC aggregate collection—85 percent—are monographs. The second largest category of materials—12 percent—is comprised of the "collections" discussed above: e.g., archival materials, vertical files, and so on. Serials

accounted for about 2 percent of the titles in the aggregate collection, and the remaining 1 percent included a variety of other, sparsely represented materials types, such as integrating resources, monographic component parts, and serial component parts.

Languages

More than 150 different languages were represented among the materials in the NYARC aggregate collection. Not surprisingly, English-language materials predominate, accounting for 49 percent of the titles in the aggregate collection. French was the next most common language, at 14 percent, followed by German (11 percent); Italian (7 percent); and Spanish (3 percent). Although roughly half of the titles in the NYARC collection are English, and 84 percent are distributed across only five languages, there is nevertheless a great deal of language diversity to be found in the aggregate holdings of the four NYARC institutions, at least in terms of representation if not quantity.

Publication dates

Another dimension along which to sketch the contours of the NYARC aggregate collection is the distribution of titles by publication date. The distribution of titles in the NYARC collection by publication date is shown in figure 3. Approximately half of the NYARC collection was published after 1970; almost a quarter was published after 1990. In comparison, the analysis of the aggregate collection of the original five libraries participating in the Google Book Search project referenced earlier revealed that about half the collection was published after 1974; another study which examined the system-wide aggregate collection of print books[4] found that about half of this collection was published after 1977. The NYARC collection therefore exhibits a median "age" similar to those associated with the "Google 5" and system-wide aggregate collections.

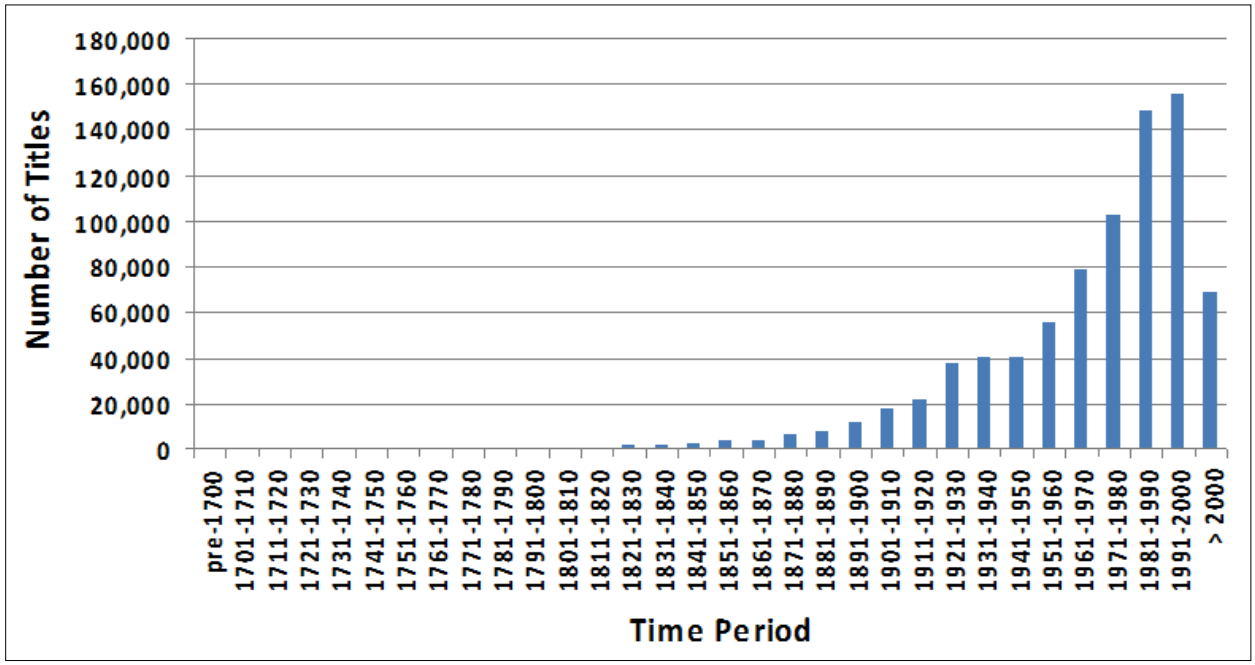


Figure 3. Distribution of publication dates in the NYARC aggregate collection

While a substantial fraction of the NYARC collection is of fairly recent publication, a significant number of titles are much older. About 9 percent were published prior to 1923, which can be interpreted as a rough demarcation between in- and out-of-copyright materials. Determination of the copyright status of the titles in the NYARC collection could be critical information in the context of activities such as digitization or other forms of repurposing. The analysis of the NYARC collection's publication patterns indicates that a little more than 90,000 titles are in the public domain, presumably with no copyright restrictions attached. Assuming that this out-of-copyright material is of general interest to the art community, it could be a strong candidate for digitization and online access.

Exhibition catalogs and auction catalogs

Exhibition catalogs and auction catalogs are items of special interest to art museum libraries; consequently, our analysis of the NYARC aggregate collection paid particular attention to these materials. Exhibition catalogs are publications in conjunction with an exhibit, often including images of objects on display, as well as essays documenting the show. They are valuable as a form of "permanent record" of an otherwise transitory event, and might also serve as a means of documenting the intellectual effort involved in selecting or arranging the pieces included in the exhibition. Auction catalogs are listings of objects available for bidding at an auction, and are important records to establish the provenance and historic valuation of a particular item.

More than 250,000 unique exhibition catalog titles can be found in the NYARC aggregate collection, along with more than 130,000 unique auction catalog titles. Exhibition catalogs account for 26 percent of the NYARC aggregate collection; auction catalogs account for 14 percent. Taken together, both types of catalog account for 40 percent of the combined holdings of the four NYARC institutions. Clearly, extensive holdings of exhibition and auction catalogs represent one of the distinctive features of an art museum library, and a core aspect of its collecting activity.

Focusing on exhibition catalogs for the moment, three of the individual NYARC collections devote roughly a quarter of their holdings to this class of material. The exhibition catalog holdings of the fourth NYARC institution, however, account for 40 percent of its collection. The NYARC libraries were particularly interested in a pair-wise analysis of exhibition catalog holdings overlap—that is, the degree to which the exhibition catalog holdings of any given pair of NYARC libraries coincided. Results ranged from a low of 6 percent to a high of 13 percent, where the percentages are interpreted as the fraction of the two institutions' combined exhibition catalog holdings held by both institutions. These results suggest that exhibition catalog holdings are fairly unique across the NYARC institutions. The degree of overlap, however, is not insignificant, and suggests that exhibition catalogs might be one area where the NYARC members could collaborate to reduce redundant collecting activity.

Auction catalogs are another important part of the NYARC aggregate collection. Collectively, the NYARC libraries hold more than 130,000 unique auction catalog titles, accounting for about 14 percent of the aggregate collection. Closer inspection of the NYARC auction catalog holdings suggests that more than 2,800 distinct auction houses are represented in the collective NYARC holdings. Examination of the individual collecting patterns for auction catalogs among the NYARC members shows a substantial amount of overlap in terms of the auction houses that are the focus of collecting activity, especially in regard to the two NYARC libraries with the largest auction catalog collections. Given the vast number of auction houses whose catalogs are of interest to the art community, combined with the evidence that there is already significant overlap across the NYARC members in terms of coverage of many of these auction houses, opportunities may exist to optimize collecting activity for these materials within the NYARC framework of cooperation. For example, there may be mutually beneficial arrangements in which the auction catalog collecting activity is apportioned across the NYARC members in such a way as to maximize coverage of auction houses while minimizing redundant collecting activity.

IV. Beyond the NYARC aggregate collection

The analysis to this point has examined holdings patterns and the degree of uniqueness within the NYARC aggregate collection. But how does the NYARC collection itself compare to a wider

world? How unique is the NYARC collection vis-à-vis the collections of other libraries? A variety of further comparative studies provide a sense of how the NYARC aggregate holdings compare to other collections, such as the system-wide collection of libraries as approximated by the RUC/SCIPIO and, to an even greater extent, the holdings in WorldCat. Comparisons to the collections of other New York-area research institution, as well as the collections of a peer institutions round out the picture.

NYARC vis-à-vis RUC and SCIPIO

To assess the NYARC aggregate collection in comparison to the RUC and SCIPIO, we examined "cluster sizes" for each NYARC title. In the RUC and SCIPIO environments, each title is associated with a cluster of records, with each record corresponding to an institution holding the title in its collection. The size of the cluster, therefore, indicates the number of institutions holding the title, at least in terms of those institutions whose holdings are represented in the RUC and SCIPIO databases.

To conduct this analysis, RUC and SCIPIO titles were segregated and analyzed separately; since SCIPIO represents a class of materials (auction catalogs) of special interest to art museum libraries, it is likely that cluster sizes will tend to be smaller on average than those associated with other materials. Results for the RUC titles are reported in figure 4. Examination of the RUC titles in the NYARC collection yields 319,684 titles, or 33 percent, with clusters of size equal to one (highlighted in red in figure 4), indicating that at least in the context of the holdings represented in the RUC database, a NYARC institution is the only institution holding the title. This suggests that a large proportion of the NYARC aggregate collection exhibits a degree of uniqueness that extends beyond the limited context of the NYARC institutions themselves.

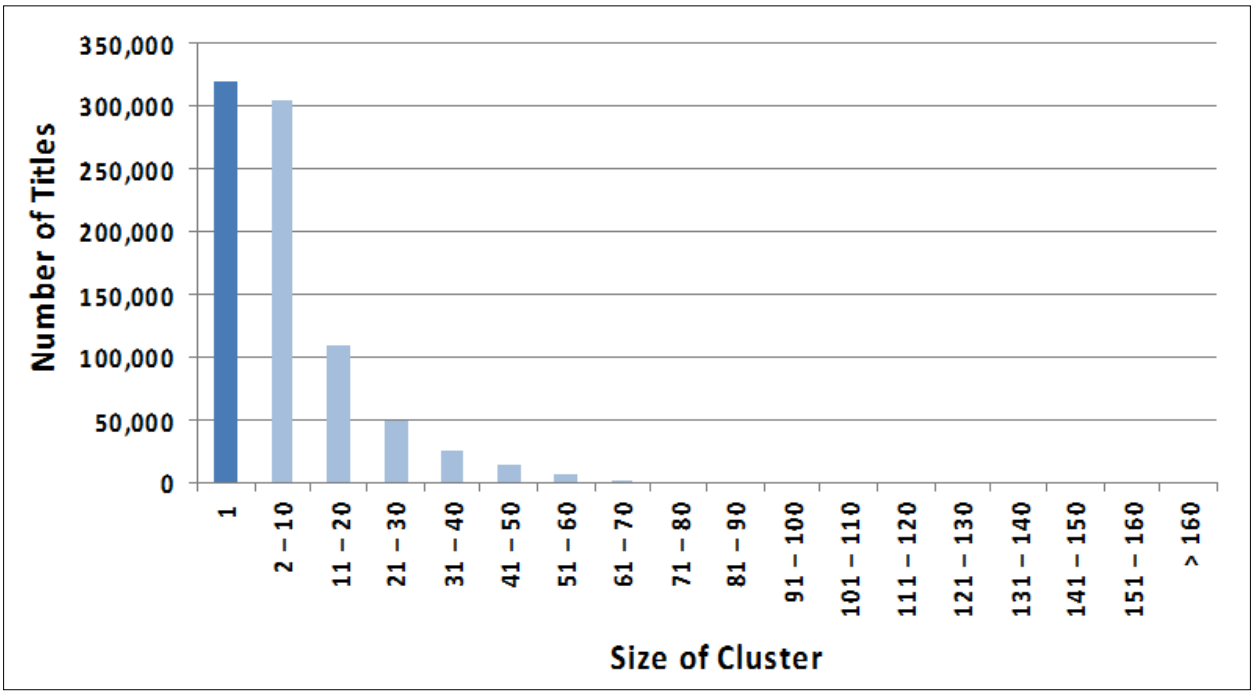


Figure 4. Distribution of RUC cluster sizes for titles in NYARC aggregate collection

Similar results were obtained for the auction catalogs in the SCIPIO database (shown in figure 5), where 30,077 titles were in clusters of size equal to one, indicating that at least in the context of the institutions whose holdings are represented in SCIPIO, the title is held exclusively by a single NYARC institution. Taken together, these results suggest that the NYARC aggregate collection represents a highly unique resource even when examined within the wider scope of all institutions with holdings represented in the RUC and SCIPIO databases.

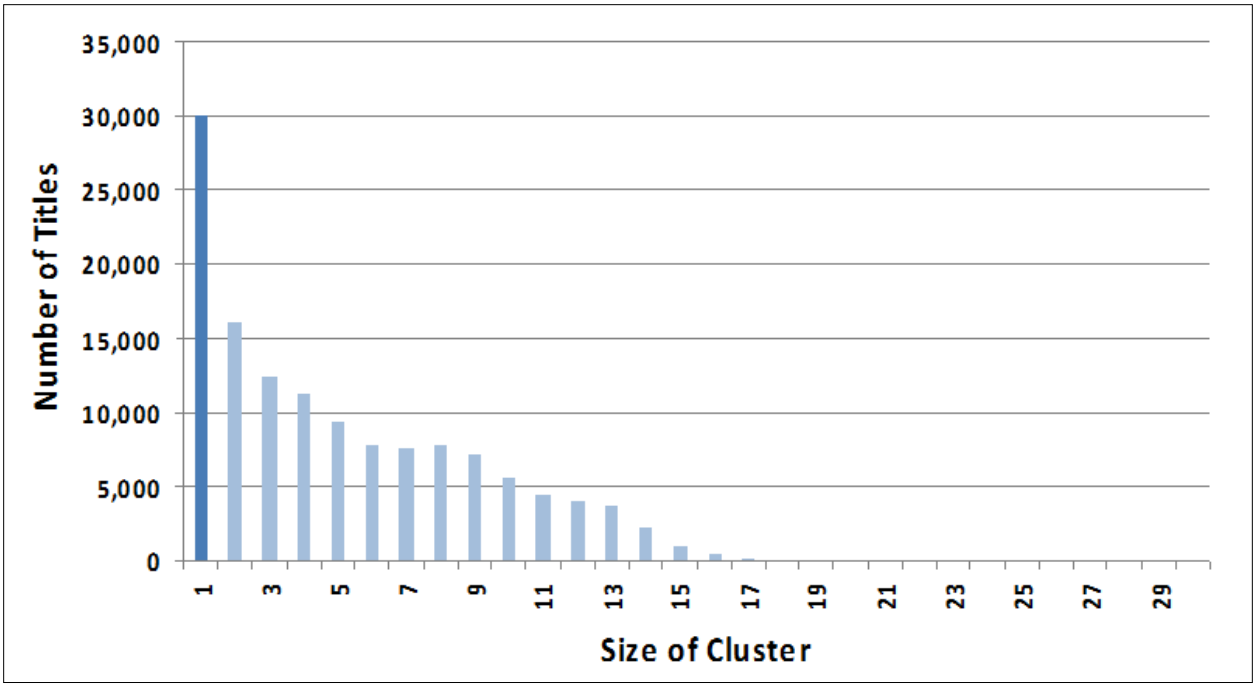


Figure 5. Distribution of SCIPIO cluster sizes for titles in NYARC aggregate collection

Finally, we examined the degree of uniqueness associated with the NYARC institutions and their collections by computing the percentage of each individual NYARC collection that was unique compared to the rest of the holdings represented in the RUC and SCIPIO databases. This analysis yielded results that varied significantly over the four NYARC institutions: roughly half of one of the NYARC collections was unique compared to the rest of RUC and SCIPIO; at the other extreme, only 13 percent of another NYARC collection was unique compared to other RUC and SCIPIO holdings. The other two NYARC institutions registered about a third of their collections as unique vis-à-vis the rest of RUC and SCIPIO.

NYARC vis-à-vis WorldCat

The second external comparison was also the largest in scale: the NYARC aggregate collection was compared to WorldCat, the world's largest bibliographic database. Since WorldCat embodies the combined holdings of thousands of libraries world-wide, it arguably serves as the most accurate proxy for the "system-wide library collection"—the aggregate collection of libraries everywhere. The NYARC libraries were interested in knowing to what degree their aggregate holdings overlapped with those of the general library community. Put another way, is the NYARC aggregate resource easily replicable by aggregating the collections of some other combination of libraries?

Auction catalogs were excluded from the analysis, since this category of material, while commonly held by art libraries, is not usually collected by other institutions, and therefore would tend to inflate the degree of uniqueness associated with the NYARC aggregate collection. Exclusion of auction catalogs left approximately 830,000 unique titles in the NYARC collection. When compared against WorldCat, approximately 60 percent of these titles were held by at least one other library with holdings represented in WorldCat. Alternatively, about 40 percent of the NYARC aggregate collection constitutes a unique resource in the general library environment as represented by WorldCat, a slightly higher percentage compared to the 33 percent from the RUC comparison. Both figures, however, confirm that if a user wishes to consult a title in this portion of the NYARC aggregate collection, it is likely that the only option will be to do so through one of the NYARC libraries.

This result suggests that the NYARC institutions hold a significant amount of material that is, at the least, not easily obtainable from other institutions. Given this, it follows that any measure that improves the general accessibility of the NYARC collective collection holds the promise of significant benefits in terms of supporting the research and learning needs of an art community that extends well beyond the on-site visitors to the four NYARC institutions. These benefits spring from the relatively unique niche the NYARC aggregate collection occupies in the library landscape.

NYARC vis-à-vis NYPL, Columbia and NYU

The NYARC libraries were also interested in determining the degree to which their aggregate holdings overlapped with other, non-art museum research libraries in the New York City area. To make this comparison, the NYARC aggregate collection was compared against the collective holdings of the libraries at New York University and Columbia University, as well as the New York Public Library. Results of this comparison revealed that about a third of the NYARC aggregate collection was held at one or more of these New York research institutions; two-thirds of the NYARC collection, on the other hand, was not. As with the comparison to WorldCat, these results once again speak to the uniqueness of the NYARC "collective collection," but this time in relation to a smaller "space": the landscape of information resources available in the New York City area. Several interesting possibilities emerge from this analysis. The fact that as much as a third of the NYARC aggregate collection overlaps with resources available at other, geographically proximate institutions suggests a possible opportunity for the NYARC institutions to relinquish some of their collecting activities to these institutions, and re-allocate them toward the areas of distinctiveness of their collections. On the other hand, the fact that two-thirds of the NYARC collection is unique relative to the collections at NYU, Columbia, and the NYPL suggests that users at these latter institutions might benefit greatly from easy access to a unique art resource in close proximity to their primary research centers. A reciprocal agreement would also benefit NYARC patrons,

including curators. As scholarship becomes more and more interdisciplinary, the materials available in other local research libraries might provide a valuable backdrop for the highly specialized NYARC content. In summary, opportunities might exist for the NYARC libraries to reduce their collecting activities in areas characterized by high redundancy with other nearby institutions, strengthen the distinctive or unique aspects of their collective holdings, and enhance the accessibility of the aggregate NYARC resource to nearby researchers and students while at the same time expanding available materials for their own audiences.

NYARC vis-à-vis the Getty Research Institute

Finally, the NYARC institutions were interested in a comparison between their aggregate holdings and those of a non-NYARC peer institution—that is, the holdings of another art library. To shed some light on this question, the NYARC aggregate collection was compared to the holdings of the Getty Research Institute in Los Angeles, California. Analysis revealed that the holdings overlap between NYARC and the Research Institute was about 20 percent—in other words, 20 percent of the materials in the NYARC aggregate collection were also available at the Getty. This degree of overlap falls below the corresponding percentages for the system-wide library collection (WorldCat) and the three New York research institutions. In a sense, however, a collection overlap of 20 percent seems proportionately high: in this case, the comparison is between the NYARC collection and the holdings of a single institution. In contrast, the comparison to the New York research centers involved the holdings of three libraries, while the NYARC-WorldCat comparison involved thousands of libraries. If the overlap with the Getty Research Institute is indeed proportionately high, this can be at least partially attributed to the fact that the Getty is a peer institution, and like the NYARC members, specializes in collecting in art-related subject areas.

Whether or not one considers the NYARC-Getty overlap high, the fact remains that 80 percent of the NYARC aggregate collection is not available at the Getty. Once again, uniqueness seems to be the prevailing theme. In this case, the uniqueness is manifested in the “space” of peer institutions, and even here, the NYARC collection stands out as a unique art resource.

V. Possible applications of the analysis

The analysis reported in the previous three sections paints a general picture of the NYARC aggregate collection: size, holdings patterns, characteristics of its content, and uniqueness vis-à-vis other collections. Information of this kind is useful as a descriptive tool, but its true value is released when it is actionable: that is, when it can be directly applied to a range of decision-making needs. In what areas could this aggregate collection analysis be applied within the NYARC framework of cooperation? After sharing the results of the analysis with the NYARC institutions, a teleconference was held with representatives from the four libraries to

discuss decision-making areas in which they felt the analysis might be particularly illuminating. As the discussion proceeded, four major areas emerged; these are listed below, accompanied by examples of aggregate collection analysis particularly relevant to each area.

- Shared storage: reduce cost and leverage economies of scale through collaborative print storage solutions
    - *Identify print materials held by multiple NYARC libraries*
- Resource sharing: expand the landscape of information resources available to users, regardless of location
    - *Identify patterns and concentrations of holdings in various subject areas across the NYARC institutions*
- Digitization: improve access to rare or unique materials through digital surrogates
    - *Identify uniquely held or rare materials at each NYARC institution*
- Partnerships with other libraries: establish cooperative arrangements with local and peer institutions in areas like collection development and reciprocal borrowing agreements
    - *Assess strengths and weaknesses of NYARC collection vis-à-vis collections held by other institutions or groups of institutions*

Knowledge of the contours of the NYARC collective collection provides a foundation for deeper forms of collaboration in all of these areas. NYARC is but one example of the increasing importance of aggregate collections, spurred by the growing "interconnectedness" among libraries as networks of cooperation within the library community develop and expand. As the examples listed above suggest, the opportunities for creating value through collective action, or by aligning local collections with certain aspects of a larger context, are numerous and diverse. Aggregate collection analysis illuminates these opportunities, and aids the formulation of appropriate decisions and policies to act on them.

Knowledge of the contours of the NYARC collective collection provides a foundation for deeper forms of collaboration in all of these areas. NYARC is but one example of the increasing importance of aggregate collections, spurred by the growing "interconnectedness" among libraries as networks of cooperation within the library community develop and expand. As the examples listed above suggest, the opportunities for creating value through collective action, or by aligning local collections with certain aspects of a larger context, are numerous and diverse. Aggregate collection analysis illuminates these opportunities, and aids the formulation of appropriate decisions and policies to act on them.

VI. Conclusion

As the analysis of the collective collection of four New York City-area art museum libraries demonstrates, studies of aggregate collections provide valuable intelligence in support of collaborative initiatives impacting multiple institutions and their collections. Awareness of broader contexts extending beyond the boundaries of the local collection is becoming increasingly important for libraries and other collecting institutions. Networks of collaborating institutions are growing in areas such as mass digitization, cooperative print storage, collection development, and shared discovery environments. As these networks continue to develop and expand, the need for aggregate collection analysis will grow commensurately. Aggregate collection analysis facilitates collective action on the part of multiple institutions, and even informs local decision-making by placing it against a wider context.

Sketching out the contours of the NYARC aggregate collection supports collaboration among the four libraries across a variety of dimensions: eliminating redundant collecting effort; identifying and leveraging individual institutional strengths within a framework of cooperation; and as three of the NYARC members have done, consolidating their collective holdings into an integrated discovery environment, thus creating a collective art resource of considerable proportions to which the art community will naturally gravitate. Aggregate collection analysis of the kind reported in this paper provides a useful context against which discussions of possible future collaborations can take place.

There are a variety of ways to conceive of aggregating library collections: aggregation by geography, aggregation by subject specialty, aggregation by consortial affiliation, and so on. The combined collection of the NYARC institutions represents what is perhaps a rare breed of aggregation, in that their collective holdings touch on all of these dimensions: they are clustered in a fairly narrow geographical area; they are all art-centric collections; and they represent the holdings of institutions who are members of a consortium formed to explore collaborative opportunities. In short, the NYARC institutions are clustered together in a variety of spaces—geography, subject, affiliation—and by extension, their collections are clustered in these spaces as well. Consequently, the incentives to analyze the scope and characteristics of the collective NYARC holdings are clear. Looking beyond the NYARC members to the general library community, the opportunities for aggregate collection analysis might not always be this apparent, but they nevertheless exist and can be leveraged to make cooperation and collective action among libraries as fruitful as possible.

Notes and References

1 The news release is available at http://www.iii.com/news/pr.php.

2 NYARC participants noted some materials in their collections that had not yet been cataloged and/or loaded into the RUC or SCIPIO at the time of the extraction: 25,000 and 32,000 vertical files respectively from two NYARC institutions, and 5,000 and 15,000 auction catalogs respectively from two institutions. These materials are not included in the analysis. Auction catalogs from one of the NYARC institutions were only available in RUC, not in SCIPIO; these were not included in the analysis of auction catalogs reported later in the paper. It should be noted that identification of unique titles in the analysis is based on clustering within the RUC and SCIPIO databases. Variations in cataloging may introduce a small margin of error if they prevent identical titles from clustering.

3 Lavoie, Brian, Lynn Silipigni Connaway, and Lorcan Dempsey. 2005. "Anatomy of Aggregate Collections: The Example of Google Print for Libraries." *D-Lib Magazine*, 11(9) (September). Available at: http://www.dlib.org/dlib/september05/lavoie/09lavoie.html.

4 Schonfeld, Roger C. and Brian F. Lavoie. 2006. "Books without Boundaries: A Brief Tour of the System-wide Print Book Collection." *Journal of Electronic Publishing*. 9(2) (Summer). http://hdl.handle.net/2027/spo.3336451.0009.208.

Print Management at "Mega-scale": A Regional Perspective on Print Book Collections in North America

Brian Lavoie, Constance Malpas, JD Shipengrover

Acknowledgments

We wish to thank Michelle Alexopoulos, Ivy Anderson, James Bunnelle, Lorcan Dempsey, David Lewis, Rick Lugg, Lars Meyer, Roger Schonfeld, Emily Stambaugh, and Thomas Teper for their thoughtful comments on a draft version of this report; their feedback was immensely helpful in improving the final version. We also thank Michelle Alexopoulos for her aid in obtaining the ZIP/postal code data used to construct the mega-regional collections analyzed in the report. We owe debts of gratitude to several OCLC colleagues: Bruce Washburn, for his assistance in producing the HathiTrust overlap findings; and Lorcan Dempsey, to whom the credit belongs for perceiving the mega-regions framework as a valuable context for exploring library data, and who encouraged us to find application for the framework in our work.

Introduction

The future of print book collections has received much attention, as libraries consider strategies to manage down print while transitioning to digital alternatives. The opportunity for collaboration is a recurring theme in these discussions. The OCLC Research report *Cloud-sourcing Research Collections: Managing Print in the Mass-digitized Library Environment* (Malpas 2011) considers the prospects for shifting the locus of print book management models from local collections to regionally consolidated shared collections, and concludes that while the necessary policy and technical infrastructures have yet to be developed, a "system-wide reorganization of collections and services that maximize the business value of print as a cooperative resource is both feasible and capable of producing great benefit to the academic library community" (p. 64).

This paper was originally published in July 2012 as an OCLC Research report.
http://www.oclc.org/research/publications/library/2012/2012-05.pdf.

As the *Cloud-sourcing* report acknowledges, much work remains to be done before a system of consolidated regional print collections becomes a reality. Nevertheless, it is interesting to speculate on an imagined future where such a system has materialized. A key question is the nature of the consolidated regional collections themselves—what would they look like? How similar or dissimilar would they be? Taken together, would the regional collections constitute a system of similar print book aggregations duplicated in different geographical regions, or would each collection represent a relatively unique component of the broader, system-wide print book corpus? These and other questions are relevant to a variety of broader issues, including mass digitization, resource sharing, and preservation.

The answers depend on how the collections are consolidated, or in other words, how the regions are defined. Several regional models for shared print book storage facilities are in evidence today. For example, the Five College Library Depository is shared by Amherst, Hampshire, Mount Holyoke, and Smith Colleges, and the University of Massachusetts Amherst. All of these institutions are clustered in the Connecticut River Valley in western Massachusetts. On a larger scale, the Northern and Southern Regional Library Facilities provide book storage capacity for the northern and southern campuses, respectively, of the University of California system. And on an even larger scale, the Western Regional Storage Trust (WEST) project proposes a distributed print repository service serving research libraries in the western United States.

Investigating the characteristics of a system of regionally-consolidated shared print book collections requires two elements: a model of regional consolidation, and data to support analysis of collections within that framework. This paper employs the mega-regions framework for the first and the WorldCat bibliographic database for the second. Mega-regions are geographical regions defined on the basis of economic integration and other forms of interdependence. The mega-regions framework has the benefit of basing consolidation on a substantive underpinning of shared traditions, mutual interests, and the needs of an overlapping constituency.

This report explores a counterfactual scenario where local US and Canadian print book collections are consolidated into regional shared collections based on the mega-regions framework. We begin by briefly reviewing the conclusions from the *Cloud-sourcing* report, and then present a simple framework that organizes the landscape of print book collection consolidation models and distinguishes the basic assumptions underpinning the *Cloud-sourcing* report and the present report. We then introduce the mega-regions framework, and use WorldCat data to construct twelve mega-regional consolidated print book collections. Analysis of the regional collections is synthesized into a set of stylized facts describing their salient characteristics, as well as key cross-regional relationships among the collections. The

stylized facts motivate a number of key implications regarding access, management, preservation, and other topics considered in the context of a network of regionally consolidated print book collections.

Context

The analysis in this paper builds upon findings from the *Cloud-sourcing* report, which was motivated by a growing concern within the academic library community about the perceived decline in use (measured by circulation) of print collections, as well as the anticipated shift toward use of, not to say preference for, digital surrogates produced through mass-digitization programs. The report addressed these issues by investigating the overlap across print book collections in US academic libraries and the growing corpus of digitized books. Given that few (if any) library directors would withdraw a local print book collection in favor of digital surrogates without a guarantee of continued access to print originals, and in view of the cost-efficiencies of shared library storage, the report also measured the level of duplication between digitized books and physical inventory in existing shared repositories.

Several key findings emerged from this investigation. First, a significant share of the print book collections in Association of Research Libraries (ARL) institutions is duplicated in the HathiTrust Digital Library digitized book corpus; moreover, the rate of duplication showed a steady growth over a twelve-month period. The median level of duplication[1] was about 19 percent in June 2009, and exceeded 30 percent a year later. Estimates projected the median overlap with HathiTrust to reach 36 percent by June 2011.[2] While this analysis does not take into account issues concerning the substitutability of digital surrogates for print originals, it does demonstrate that the content in HathiTrust substantially duplicates—by as much as a third or more—the print content managed at much greater expense in local ARL print collections.

Another finding was that the locally held print content duplicated in the HathiTrust library is typically held by many libraries. In other words, much of this content is neither obviously "at risk" from a preservation point of view, nor in short supply from a fulfillment perspective. Consequently, the operational concerns associated with shifting print management and access operations to a trusted partner are relatively modest. Once an acceptable digital access and use platform emerges, many academic institutions will likely seek to externalize or "outsource" their traditional print repository functions to other providers. A risk inherent in a large-scale transformation of the system-wide print book collection is that a disorderly transition from local to group management may exacerbate disparities in access and even jeopardize the preservation of distinctive print resources. A prime motivation for the present study was a concern that a reconfiguration of print books held by a relatively small number of institutions could have a dramatic effect on the library system as a whole.

The *Cloud-sourcing* report found a high level of overlap (about 75 percent) between the holdings of HathiTrust and a sample of holdings from the aggregate inventory of several large-scale shared print storage repositories. However, the overlap between an individual ARL university library, the sample print storage inventory, and the HathiTrust collection was surprisingly low, suggesting that bilateral agreements between individual institutions and storage repositories were unlikely to generate the kind of space and cost savings that library directors (or university administrators) are likely to seek in an outsourcing arrangement. The report considered two potential solutions to this problem. First, a cooperative agreement among existing large-scale library storage facilities might prove to be more effective in terms of collective preservation and on-demand fulfillment. Alternatively, individual storage facilities might choose to adopt a collection development policy that would be optimized for a shared print service, by deliberately accessioning resources that would be of value to many institutions in the region.

The solutions explored in the *Cloud-sourcing* report focus on print collections held in academic research libraries and assume physical consolidation of individual print collections into an above-the-institution aggregation. This paper takes an alternative approach, based on a broader view of library print collections—including those held in public libraries—and assumes that local print collections remain local, but are *virtually* consolidated at the regional level. The next section places this in the larger context of potential print consolidation models.

A Framework for Models of Print Consolidation

For the purposes of this report, *print consolidation* refers to any strategy undertaken by a group of institutions to achieve a mutual purpose by imposing some degree of integration across their local print collections. This definition is admittedly vague, because as will be seen, its two key components—"mutual purpose" and "degree of integration"—can be manifested in multiple ways. However, the definition is useful because it identifies the two fundamental dimensions along which any model of print consolidation can be characterized: *why and how* print collections are being consolidated.

Each dimension can be characterized in numerous ways, but to keep the discussion tractable, we will focus on two facets within each dimension. In terms of the first dimension (why print collections are consolidated), we identify two general goals or objectives. First, consolidation of print collections could be motivated by the desire to create a *shared back-up collection of print originals*, with end-users relying primarily or even exclusively on digitized surrogates for access.[3] Alternatively, the consolidated collection could serve as a *shared resource for use*, with the aggregated print book holdings of multiple institutions leveraged over a wider base of potential users.

In terms of the second dimension (how print collections are consolidated), we consider two general strategies for achieving consolidation. First, local collections can be physically combined into a single shared collection and housed at a centralized repository (or limited network of shared repositories). Alternatively, consolidation can be achieved *virtually*, where local print collections remain in the custody of their respective institutions, but are linked through a layer of services, such as a shared discovery environment and fulfillment system.[4]

Combining these two dimensions yields a simple framework (see figure 1) that serves the dual purpose of providing a high-level mapping of the print consolidation landscape, and orienting the analysis in this report within the spectrum of potential print consolidation models.

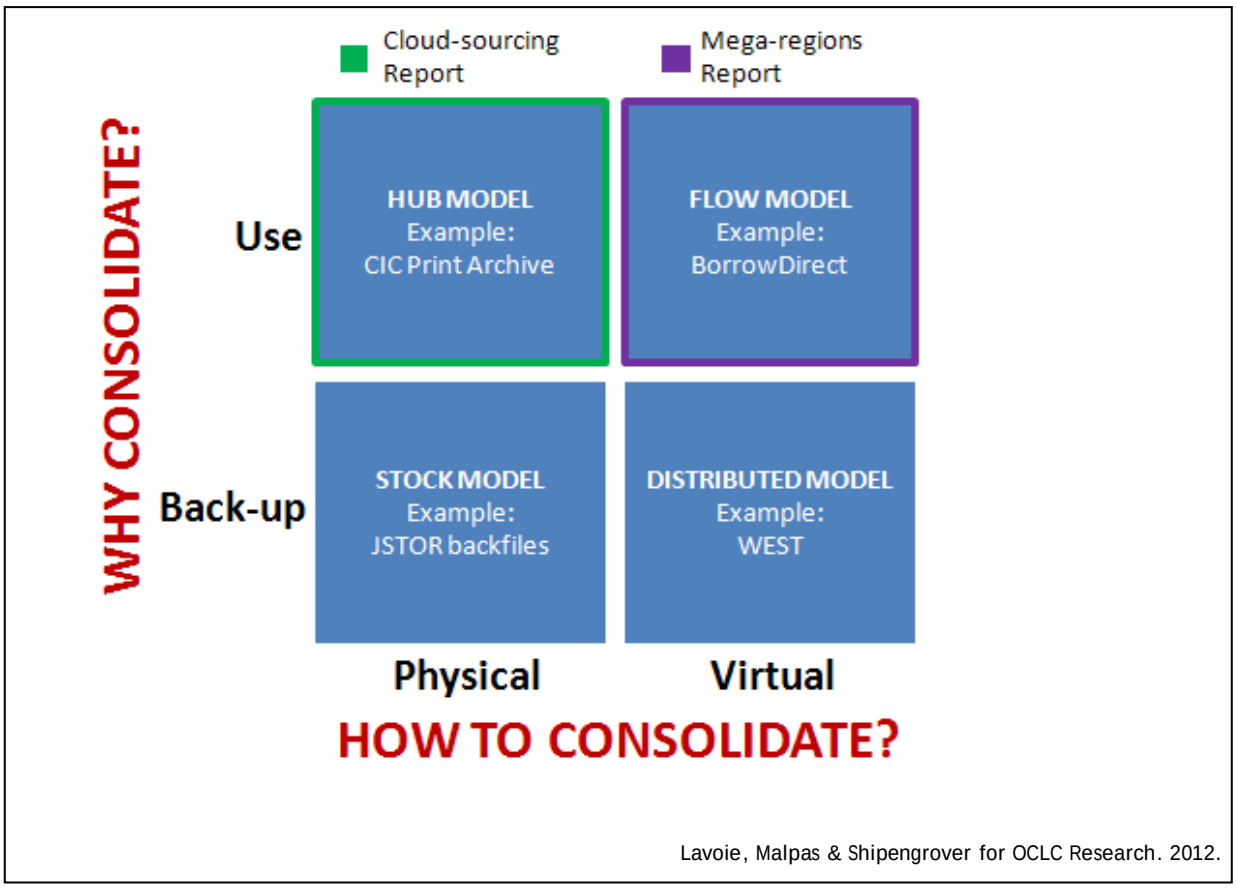


Figure 1. A framework for print collection consolidation

The framework identifies four basic models of print consolidation:

- *Hub model*: shared use of print materials is achieved through some form of physical consolidation of local collections.
- *Flow model*: shared use of print materials is achieved through some form of virtual integration across local collections.

- *Stock model*: shared back-up of print originals is achieved through a centralized consolidation of print materials into a shared repository.
- *Distributed model*: shared back-up of print originals is achieved through a virtual collection distributed across, and maintained within, local print collections.

Limiting the characterization of print consolidation models to these two dimensions omits other important aspects of consolidation. For example, these dimensions do not indicate whether local print collections are retained intact after consolidation, or some form of weeding/deduplication is implemented across the participating institutions' combined holdings; nor do they address whether future collecting activity by participating institutions is subject to cross-institutional coordination. The purpose of the framework is to identify and distinguish a set of basic models; issues such as weeding or coordination are questions that can be asked in the context of any of the models.

The framework suggests starker choices than what prevail in reality, where print consolidation strategies can shade between the various categories. The categories within each dimension are not mutually exclusive: for example, a consolidated print collection could plausibly serve as both a shared back-up *and* a shared resource.[5] Similarly, a consolidation strategy could involve some combination of a centralized repository of physically consolidated materials, supported by a network of locally managed collections.[6] However, the services and infrastructure needed to support each model are different; additionally, certain attributes of the consolidated collections themselves may align more readily to one model or the other. Given these considerations, we will treat the four models in the framework as distinct options, acknowledging that this is a simplification but still a useful conceptual device for orienting the analysis to follow.

The *Cloud-sourcing* report focused on print consolidation models falling into the upper-left hand quadrant: i.e., the hub model, where the objective of shared use is achieved through physical consolidation. In this report, we are focusing on print consolidation models represented by the upper right hand quadrant: the flow model, characterized by shared use achieved through virtual consolidation. The reason for this is two-fold. First, a recurring theme throughout current discussions of cooperative print book collection management is that institutions continue to favor direct access to print book originals over a deliberate redirection of demand to digitized surrogates. The prevailing presumption is that print books in library collections are intended to be accessed and used, rather than serve merely as back-ups. This is partly an accommodation to anticipated (and sometimes demonstrated) patron preference for print formats; it is also a pragmatic stance, given the overwhelming dominance of in-copyright titles in most library collections, as well as digitized book collections.[7] Second, there is as yet no indication that institutions are willing to dispense

entirely with their local print collections, although there is certainly strong interest in making management of print collections more efficient and less costly. Given these considerations—a focus on print books for use, and the likelihood that institutions will continue to manage print book collections locally for the foreseeable future—the flow model was chosen as the basis for the analysis in this report.

Given this, it is useful to say a little more about flow models. A flow model for the management of print collections focuses on virtually consolidating local collections into a shared resource for use, by linking them though a layer of shared services. In these circumstances, access is the primary service offering, with print materials "flowing" through the network of participating institutions to wherever needed. The chief benefit of a flow model approach to print management is the opportunity to leverage greater value from the legacy investment in print collections, by encouraging and facilitating greater use over a larger user base. This is achieved by combining a group of individual print collections into a larger and richer collective collection, which is then made available to users at all participating institutions. Attributes of the flow model are reflected in current resource sharing (ILL) networks, although such networks vary in the degree of integration across collections and access services. A well-functioning flow model helps optimize supply and demand in the collective collection by facilitating the movement of print materials from various points of supply (local collections) to the point of need (users anywhere within the network).

Distinctiveness is a desirable feature of local collections in the context of a flow model. A key benefit of a flow model approach is to expand the scope and depth of the print book offering to all users across participating institutions. If a significant portion of each participating institution's print book collection is distinctive—that is, comprised of publications not widely available at other institutions—then combining print book holdings into a collective collection yields a print book resource that is, from the perspective of the user, far more extensive than what is on hand locally. In contrast, the more similar collections are, the smaller the "gains from trade," in that access to the collective collection would offer little beyond what is available locally. Of course, substantial operational efficiencies and cost avoidance might still be achieved through some rationalization of duplicative holdings.

Since by definition flow models involve a virtual consolidation of print inventory, good data about local print book collections is essential. Consolidation occurs not at the level of the physical collections themselves, but instead within a layer of services that extends over all collections in the region and permits them to be managed and accessed as a cohesive whole. The service layer will be data-driven, and therefore its ability to present distributed print book holdings as a "regional collection" and offer functionalities operating on that

collection—such as support for cooperative collection management decision-making, or region-wide discovery and fulfillment services—will depend on the accuracy and completeness of the underlying data.

The flow model is illustrated by the Borrow Direct partnership between Brown University, Columbia University, the Center for Research Libraries, Cornell University, Dartmouth College, Harvard University, MIT, University of Pennsylvania, Princeton University, and Yale University. Borrow Direct permits faculty and students at each of the partner institutions to easily discover, request, and receive delivery of print books and other materials located at any of the other institutions. Although there are some limitations on cross-institutional borrowing privileges (e.g., one physical volume per request, loan renewal not permitted), users of Borrow Direct benefit from the larger scope and depth of the partners' collective collection, and the speed with which requested materials can be delivered to the user's location (Nitecki 2009). Each Borrow Direct institution maintains its own print collection but a layer of services link them together into a virtual collective collection. Greater value is extracted from the collective print investment by making more materials available to more users.

Mega-regions: A Framework for Consolidation

Given a model of print consolidation, a choice must be made as to the level of aggregation underpinning the consolidation. In other words, how many (and which) institutions will be involved, and where are they located? For the analysis in this report, we chose to examine consolidation at the regional level. Regions tend to be bound together by ties that can both motivate and facilitate interaction between organizations within the region, such as geographical proximity, shared infrastructure, and economic interdependencies. These ties are well-suited to support a print consolidation model based on virtual consolidation and flows of materials around the system. The logistics of supporting a flow model of print consolidation would likely be simpler and more efficient within a region, in comparison to a grouping of geographically dispersed and disconnected institutions. Moreover, regions seem to be a natural scale of aggregation for print consolidation. Regional clusters of cooperative activity seem to be where current print management initiatives are gravitating: many discussions regarding cooperative print management are organized at the regional level, sometimes involving established regional consortia. For example, a recent *Chronicle of Higher Education* article notes that the WEST project aims to build a "large-scale regional trust for print journal archives," while "talks are under way about setting up similar regional repositories in the Northeast and Southeast" (Howard 2011).

"Region" is a nebulous term, and can be defined at a variety of scales. We operationalize the concept of a region by adopting the mega-regions framework described by Richard Florida, Tim Gulden, and Charlotta Mellander in the 2008 paper, *The Rise of the Mega-region* (see also

Florida 2008). A mega-region is a geographical concentration of population and economic activity, generally subsuming multiple metropolitan areas and their surrounding hinterlands, and linked together through a complex connective tissue of economic interdependency, shared infrastructure, a common cultural history, and other mutual interests. Florida et al. observe that "[t]he mega-regions of today perform functions similar to those of the great cities of the past—massing together talent, productive capability, innovation and markets. But they do this on a far larger scale" (Florida, Gulden, and Mellander 2008, p. 460). In contrast to Thomas Friedman's idea that the global economy is "flattening," there are, the authors argue, "a strong set of counter-forces that lead to geographic clustering and the pushing together, so to speak, of economic activity. The mega-region … is a consequence of this clustering force" (p. 460).

Florida and his colleagues used satellite imagery capturing night-time clusters of lights around the globe to identify twelve mega-regions in the US and Canada (see figure 2). "… [T]he mega-region," the researchers note, "has emerged as the new "natural" economic unit. The mega-region is not an artifact of artificial political boundaries, like the nation state or even its provinces, but the product of concentrations of centres of innovation, production, and consumer markets" (p. 461).

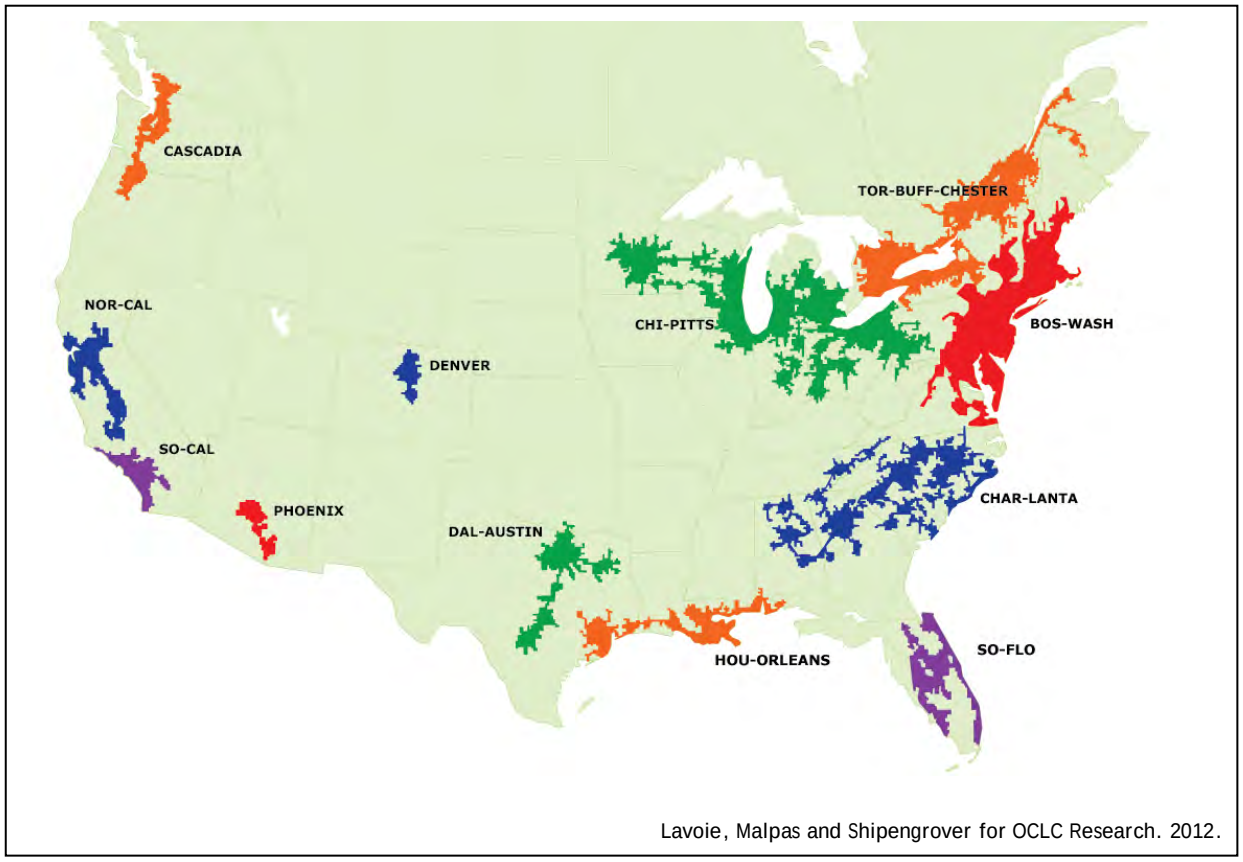


Figure 2. Mega-regions of North America[8]

As figure 2 illustrates, three of the twelve North American mega-regions extend over international boundaries: CASCADIA, CHI-PITTS, and TOR-BUFF-CHESTER. The extent of a mega-region is not limited by political boundaries, but rather by economic and cultural interdependency and mutual interests, which can occur in population centers that straddle an international border—Detroit and Windsor, for example.

Florida and his colleagues identify one mega-region in Mexico, centered around the Mexico City area. While Mexico is also part of North America, we exclude the Mexican mega-region from our analysis, and focus our attention on the remaining twelve US-Canadian mega-regions. The reason is that coverage of Mexican institutions in WorldCat is less extensive than for American and Canadian institutions, and therefore it is not clear that the Mexican presence in WorldCat would be sufficiently representative of the actual Mexican print book collection. For the remainder of the report, references to "North America" should be interpreted to mean the US and Canada only.

Mega-regions offer a compelling framework within which to think about a regional consolidation of print book collections organized as a flow model—that is, a virtual consolidation of local collections aimed at encouraging a flow of materials around the region.

Mega-regions encompass existing networks—both physical and virtual—of integration and mutual interest that could potentially absorb and support a new network of cooperative print management and shared use. As we will show below, the vast majority of the overall North American print book collection is clustered within the twelve mega-regions. In this sense, mega-regions might be a "natural unit of analysis" for cooperative print management, as well as other cooperative library activities. Finally, mega-regions represent clusters of activity—research, innovation, learning, arts, and commerce—that library collections support. Therefore, it is useful to align clusters of library resources with clusters of activities that make use of these resources.

In a sense, the North American mega-regions illustrated in figure 2 are a snapshot, in that mega-regions are not static entities but instead grow and change over time. The boundaries of the twelve mega-regions in figure 2 will likely evolve in ways that absorb parts of the hinterlands surrounding the regions. Moreover, new mega-regions may form in areas where growing economic integration and other factors serve to bind people, institutions, and activities more closely than before. These dynamics will be at work not only in mega-regions, but almost any regional framework. From the standpoint of cooperative print management, the key implication is that regional boundaries will be in flux, likely resulting in the periodic appearance of new partners and an attendant need to adjust regional cooperative arrangements.

While the mega-regions framework is a useful and convenient tool for illustrating and analyzing regional consolidation of print collections, we are not necessarily advocating mega-regions as the appropriate scale for achieving consolidation and cooperative management in practice. Assuming that regions are in fact the natural unit of consolidation, the scale at which regions are defined will depend on a host of factors, including but not limited to the location of logistical networks, existing cooperative structures and agreements, and political jurisdictions (e.g., state or provincial boundaries). Mega-regions are one of many possible forms in which regional print consolidation can be manifested; careful analysis of the alternatives will help planners arrive at the most suitable choice for their circumstances.

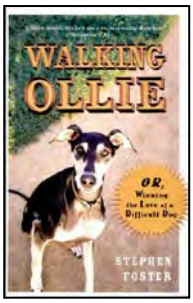
Finally, as figure 2 makes clear, there is considerable space *between* the mega-regions. We do not imply that this space is "empty" or unimportant. In fact, the space between the regions—and more specifically, the aggregation of print books located there—has interesting characteristics in its own right, with important implications for cooperative print management and shared use. We discuss the areas outside the mega-regions in detail later in the report.

Some Definitions

The following terminology is used throughout this report:

- *Print book*: a book[9] manifested in printed form. We exclude materials explicitly cataloged as theses, dissertations, or government documents from the analysis, as well as books in non-print formats such as e-books.

  Publication: a distinct edition or imprint of a work. For example, *Walking Ollie, or, Winning the Love of a Difficult Dog* is a work—a distinct intellectual creation—by the author Stephen Foster. This work has appeared as several different publications, two of which are shown below (These would be counted as two distinct print book publications in our analysis).



Foster, Stephen. 2008. *Walking Ollie, or, Winning the love of a difficult dog*. New York, N.Y.: Perigee Book.

Foster, Stephen. 2007. *Walking Ollie, or, Winning the love of a difficult dog*. London: Short.

Figure 3. Two distinct publications of the same work by Stephen Foster

- *Holding*: an indicator that a particular institution (a library or some other organization) holds at least one copy of a particular publication in its collection. Note that a holding says nothing about the number of physical copies owned by the institution, other than at least one copy is available. For example, according to its catalog, the Dallas Public Library owns three copies of the Perigree Books publication of *Walking Ollie*. All three copies would be represented in WorldCat by a single holding associated with the Dallas Public Library.[10]

- *Collective collection*: the combined holdings of a group of institutions, with duplicate holdings (i.e., those pertaining to the same publication) removed. This yields the collection of distinct publications that are held across the collections of the institutions in the group.

The North American and Mega-regional Print Book Collections

The WorldCat bibliographic database is the closest approximation available of the global collective collection—that is, the combined holdings of libraries and other institutions worldwide. While WorldCat data has certain limitations regarding coverage and interpretation of holdings information, it is nevertheless the best data source available for analysis of aggregate information resources such as regional print book collections. In January 2011, WorldCat contained 214.6 million bibliographic records representing information resources of all descriptions; these information resources accounted for nearly 1.7 billion holdings distributed across institutions all over the world.[11]

Table 1 deconstructs WorldCat into the North American print book collection.

Table 1. North American print book collection in WorldCat (January 2011)

Collection	Publications (millions)	Holdings (millions)
WorldCat	214.6	1,679.1
Print books	128.1	1,238.1
Print books in North America	45.7	889.5
Print books—US	40.9	840.0
Print books—Canada	14.2	49.4

An important caveat to note in regard to table 1, as well as other results presented in this report, is that they reflect institutional collections as they are cataloged and represented in WorldCat. The accuracy of holdings data in WorldCat may be lessened by the presence of duplicate records, cataloging errors, incomplete registration of collections, and other sources of inconsistency.

Of the 128.1 million distinct print book publications represented in WorldCat, 45.7 million are held by at least one institution located in either the US or Canada. This constitutes the *North American print book collection*, or the collective collection of print book publications held by North American institutions. Coverage of the North American collection varies considerably between the US and Canada: US institutions alone can muster 90 percent of the publications

in the North American collection, while Canadian coverage is 31 percent. Similarly, 94 percent of the holdings comprising the North American print book collection are associated with US institutions, while the remaining 6 percent are of Canadian origin.

Richard Florida and his colleagues generously provided lists of the US ZIP codes and Canadian postal codes associated with each of the twelve mega-regions defined in their 2008 paper.[12] These ZIP and postal codes were then compared to location information associated with each of the nearly 1.7 billion holdings in WorldCat. In this way, all WorldCat holdings associated with each of the twelve North American mega-regions were identified, along with all holdings located in either the US or Canada that fell outside the mega-regions. Once the holdings for a particular mega-region were identified, the subset corresponding to print book publications were extracted, and this in turn established the regional collective collection of print books. The sizes of the twelve mega-regional print book collections, measured in terms of publications and holdings, are shown in figure 4.

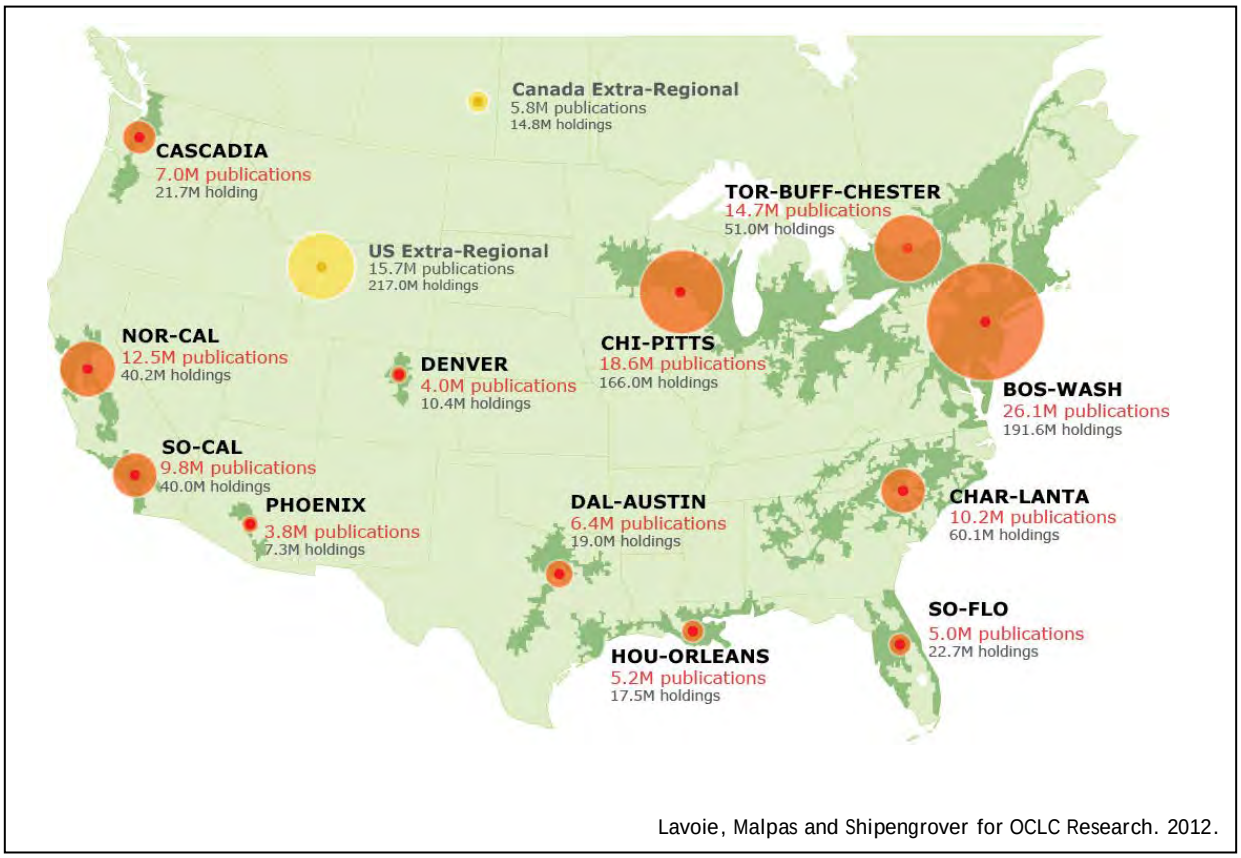


Figure 4. Sizes of the North American mega-regional print book collections. (Circles are scaled to reflect the number of print book publications in each regional collection.)

BOS-WASH is the largest regional print book collection, in terms of both distinct publications and total holdings. PHOENIX is the smallest, with only 15 percent as many publications, and 4 percent as many holdings, as BOS-WASH. The median regional collection size is 8.4 million distinct publications, and 31.3 million total holdings.

The ratio of holdings to publications provides a metric illustrating the degree to which a region's collection of distinct print book publications is "amplified" into total print book holdings around the region. Higher ratios suggest higher levels of duplication—or from an access perspective, greater levels of availability—within a region, while lower ratios suggest the opposite. Table 2 reports the holdings to publications ratio for each of the twelve regional collections.

Table 2. Holdings to publications ratio, by regional collection

Region	Holdings (millions)	Publications (millions)	Holdings/ Publication
BOS-WASH	191.6	26.1	7.34
CASCADIA	21.7	7.0	3.11
CHAR-LANTA	60.1	10.2	5.92
CHI-PITTS	146.0	18.6	8.94
DAL-AUSTIN	19.0	6.4	2.98
DENVER	10.4	4.0	2.58
HOU-ORLEANS	17.5	5.2	3.39
NOR-CAL	40.2	12.5	3.22
PHOENIX	7.3	3.8	1.91
SO-CAL	40.0	9.8	4.09
SO-FLO	22.2	5.0	4.53
TOR-BUFF-CHESTER	51.0	14.7	3.47

The holdings to publications ratio varied widely across the regions, with CHI-PITTS exhibiting the highest value (8.94), and PHOENIX the smallest (1.91). Five regions exhibit a ratio of 4.0 or higher: that is, an average of four holdings across the region per print book publication. With the exception of PHOENIX, the remaining regions all exhibit holdings to publications ratios of 2.5 or higher. These results suggest that duplication (or availability) of print book publications is, on average, relatively low within the regions: even the highest ratio, associated with CHI-PITTS, suggests that on average only about nine institutions hold a given print book publication in their collections, despite the geographical extent of the region and the many institutions it contains. We revisit this topic in more detail in the next section.

Table 3 reports coverage of the overall North American print book collection for each of the twelve regional collections.

Table 3. Regional coverage of the North American print book collection

Region	Coverage (%)
BOS-WASH	57
CASCADIA	15
CHAR-LANTA	22
CHI-PITTS	41
DAL-AUSTIN	14
DENVER	9
HOU-ORLEANS	11
NOR-CAL	27
PHOENIX	8
SO-CAL	21
SO-FLO	11
TOR-BUFF-CHESTER	32

The BOS-WASH region alone can account for nearly 60 percent of the entire North American print book collection. Other large regions, such as CHI-PITTS and TOR-BUFF-CHESTER, also exhibit significant coverage. Most regions, however, can only account for less than a quarter of the North American collection; for each of these regions, the vast majority of the print book publications available in North America are to be found elsewhere outside the region.

Before turning to a more detailed description of the twelve regional print book collections, it is useful to say a word about the areas *between* the regions. This report focuses on the regional collections, but this is not to diminish the importance of the print book holdings located outside the mega-regions. Indeed, these "extra-regional" print book holdings are significant in scale, accounting for more than 217 million holdings on 15.7 million print book publications in the US, and 14.8 million holdings on 5.8 million publications in Canada. Some of the local print book collections scattered through the extra-regional space are quite distant from even the closest mega-region; others are perched right on a mega-region's boundary, or in its nearby hinterland. Clearly, US and Canadian print book holdings located outside the mega-regions constitute an important resource, but consolidating them into collective collections, like the regional collections, can be problematic. Unlike the mega-regions, there is no obvious collaborative structure or patterns of mutual interest binding these collections together. We will say more about the US and Canadian extra-regional collections in the next section.

Stylized Facts

Mega-regions provide a framework for organizing local print book collections into regional collections. But what would these regional collections look like? To answer this question, a detailed analysis of each of the twelve mega-region print book collections was undertaken using WorldCat bibliographic and holdings data. The result was a wealth of statistics characterizing the regional collections from numerous perspectives. Rather than attempting to present all of these statistics to the reader, we instead chose to synthesize the analysis into a set of *stylized facts*—in other words, a set of broad observations based on empirical findings. Taken together, the stylized facts constitute a general description of the North American mega-region print book collections, from which a number of implications regarding access, management, and preservation can be derived. We discuss several of these implications at the end of the report.

Library operations—and reputation—are still bound up with books

The OCLC (2011) report *Perceptions of Libraries, 2010: Context and Community,* reminds us that print books continue to be synonymous with libraries and library use, noting that "[t]he library brand is 'books' … In 2005, most Americans (69%) said 'books' is the first thing that comes to mind when thinking about the library. In 2010, even more, 75%, believe that the library brand is books" (p. 38). The same report found that borrowing print books is still the top activity among library users (p. 35). Despite the attention (and funding) lavished on electronic and digital content in recent years, libraries of all types continue to devote significant resources to the management of print book collections.

While acceptance of e-books is increasing in academic and public libraries, the still-limited range of content, competing and incompatible platforms, and restrictive licensing regimes remain impediments to wide-scale adoption.[13] This has important consequences for the organization of library service provision, as well as operating expenses. As shown in a 2010 study by Paul Courant and Buzzy Nielson, the long-term costs of storing print books are significant (estimated at $4.26 per volume per year in open stacks) and relatively inelastic.[14] In contrast to the journal literature, much of which has migrated into electronic formats and aggregations managed by third-party agents, print books continue to occupy a significant share of local library space.

The long legacy of library investments in print books is reflected in the WorldCat database, where 60 percent of the bibliographic records describe print books and 75 percent of holdings are linked to print book titles. The outsized presence of print books in WorldCat records and holdings stems in part from cataloging practice. For example, title-level holdings for serials effectively mask the volume count of institutional journal holdings, which may significantly

outnumber books on a per-volume basis. Likewise, format integration (single-record cataloging of titles produced in multiple formats) means that burgeoning e-book collections are not adequately accounted for in holdings counts, since electronic holdings may be intermingled with print holdings. Yet the millions of books acquired by North American libraries over many years of operation, the shared bibliographic infrastructure created to manage them as a collective resource, and the still powerful association between the codex and the library “brand” (or stereotype) serve to highlight the importance of print books to libraries and their users.

The impact of centuries of library investment in print books can be seen at the regional level. As figure 4 illustrates, print books account for anywhere from two-thirds to three-quarters of total holdings in each of the twelve mega-regions. The same characteristic is seen across different library types. Print books account for 68 percent of ARL library collections, while non-ARL academic libraries in North America are slightly higher at 69 percent. Eighty percent of North American public library collections are print books, while North American school (K-12) library collections are even higher at 87 percent. Again, while these results must be considered in light of cataloging practice and patterns of use of WorldCat as a bibliographic utility, they are nevertheless broadly indicative, and not only illustrate the ongoing predominance of print books in library collections, but also the importance and scale of the print collection management problem. Libraries retain responsibility for managing massive amounts of print book inventory, while at the same time they are transitioning their focus—and substantial portions of their budgets—to electronic and digital collections. Moreover, libraries face economic pressures to cut costs and justify value. A new system of print book collection management is needed to accommodate these conditions.

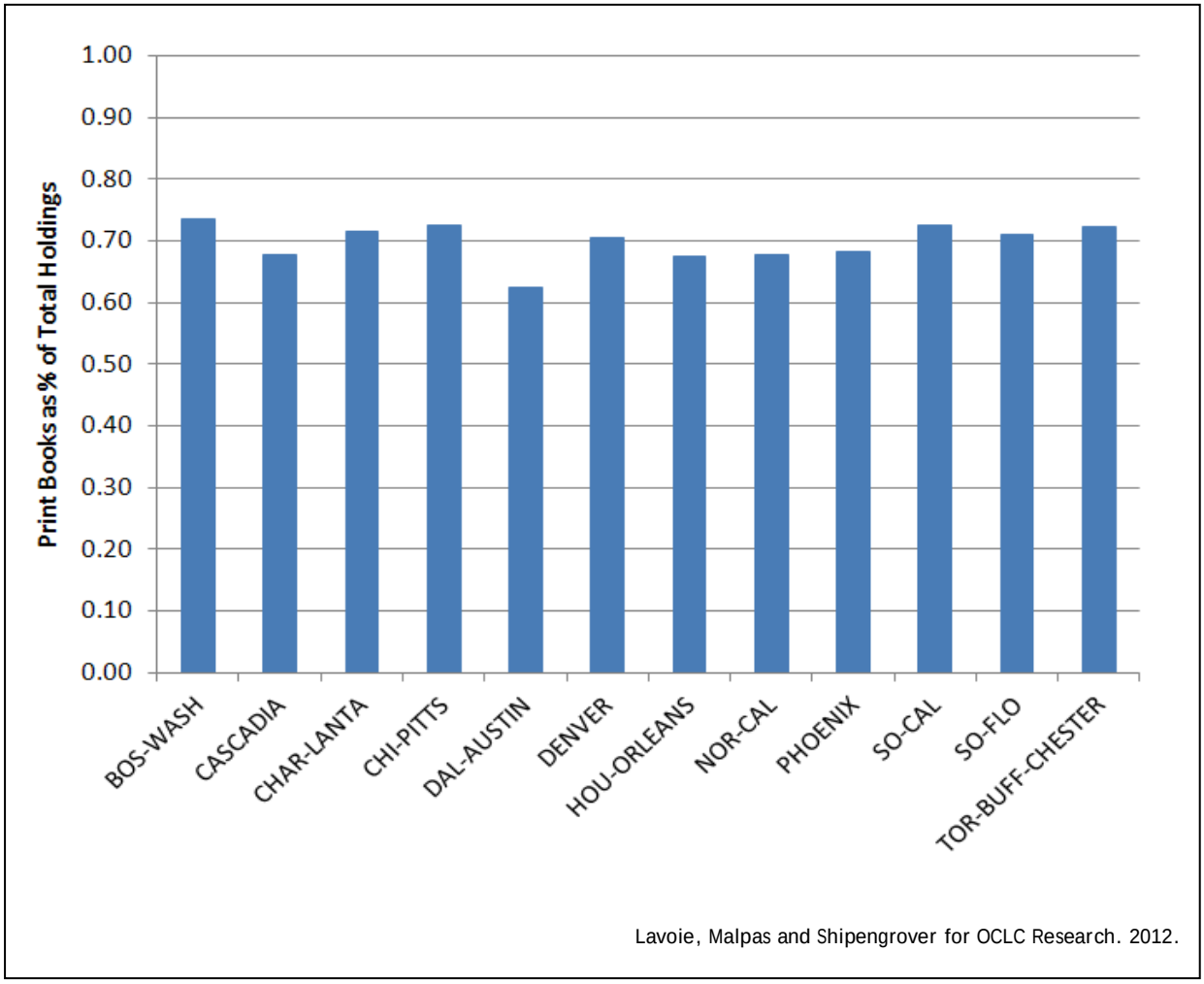


Figure 5. Print books as percent of total holdings, by mega-region

Academic institutions are the custodians of the majority of system-wide print book inventory

The success of a regionally based cooperative model of print collection management depends on engaging institutions that control significant portions of the region-wide print book inventory. As the results in figure 6 show, the majority of the print book inventory in every region is in the custody of academic institutions.

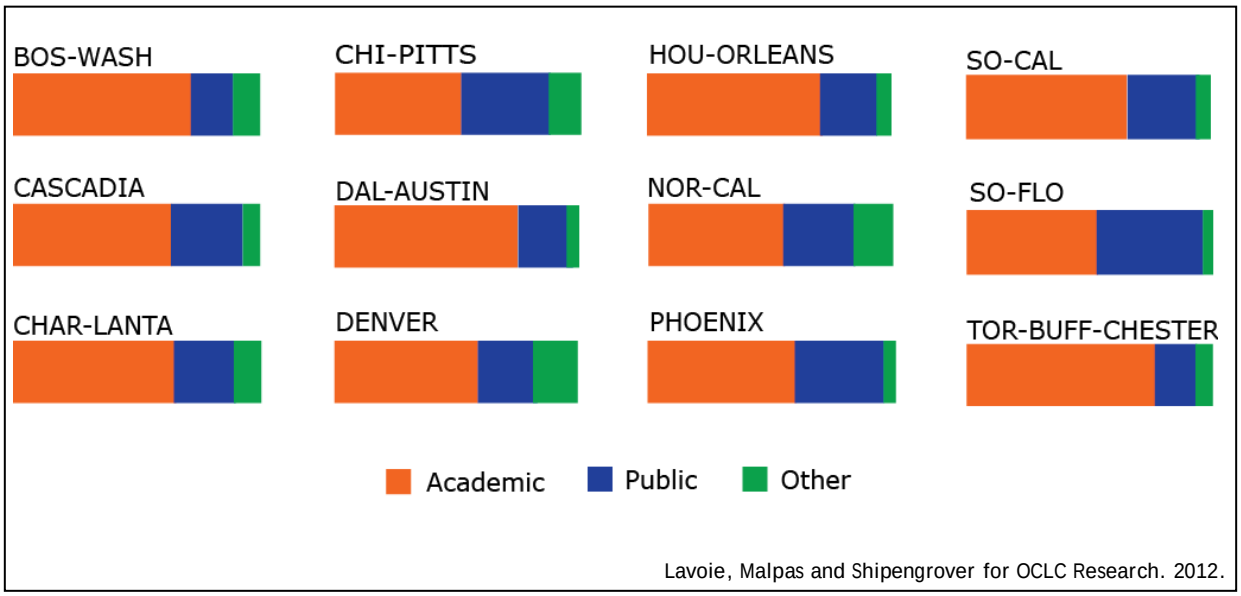


Figure 6. Share of regional print book holdings, by institution type

The extent to which academic institutions dominate print book holdings varies across regions, with the highest proportion in the TOR-BUFF-CHESTER region (76 percent), and the lowest in CHI-PITTS (51 percent). But the key point is that in every region, more than half of the regional print book inventory is in the hands of academic institutions—and in some regions, considerably more than half. We are aware that many public library holdings are not represented in WorldCat, and this will tend to amplify the relative presence of academic institutions in regional print book collections. But even taking this coverage gap into account would not serve, in our judgment, to overturn the conclusion that most print book inventory in the regional collections belongs to academic institutions, given the wide gap between the relative shares of each institution type exhibited in figure 6.

Print book holdings associated with academic institutions can be divided into those belonging to ARL institutions (the most research-intensive academic institutions), and those belonging to non-ARL academic institutions. BOS-WASH has the greatest number of print book holdings belonging to ARLs, at 65.3 million—more than twice the number of the region with the second-highest total, CHI-PITTS. However, it is in fact PHOENIX—the smallest regional collection—that has the highest percentage of its print book holdings associated with ARLs (52 percent); TOR-BUFF-CHESTER is next at 46 percent. In contrast, SO-FLO (12 percent), and CHI-PITTS and DAL-AUSTIN (both at 19 percent), are the regions with the smallest percentage of ARL holdings. Another way to assess the presence of ARLs in the regional collections is to compute the share of academic holdings in each region belonging to ARLs; figure 7 reports these results.

Considerable cross-region variation is apparent: in PHOENIX, nearly 90 percent of all academic print book holdings belong to ARLs, compared to less than a quarter in SO-FLO.

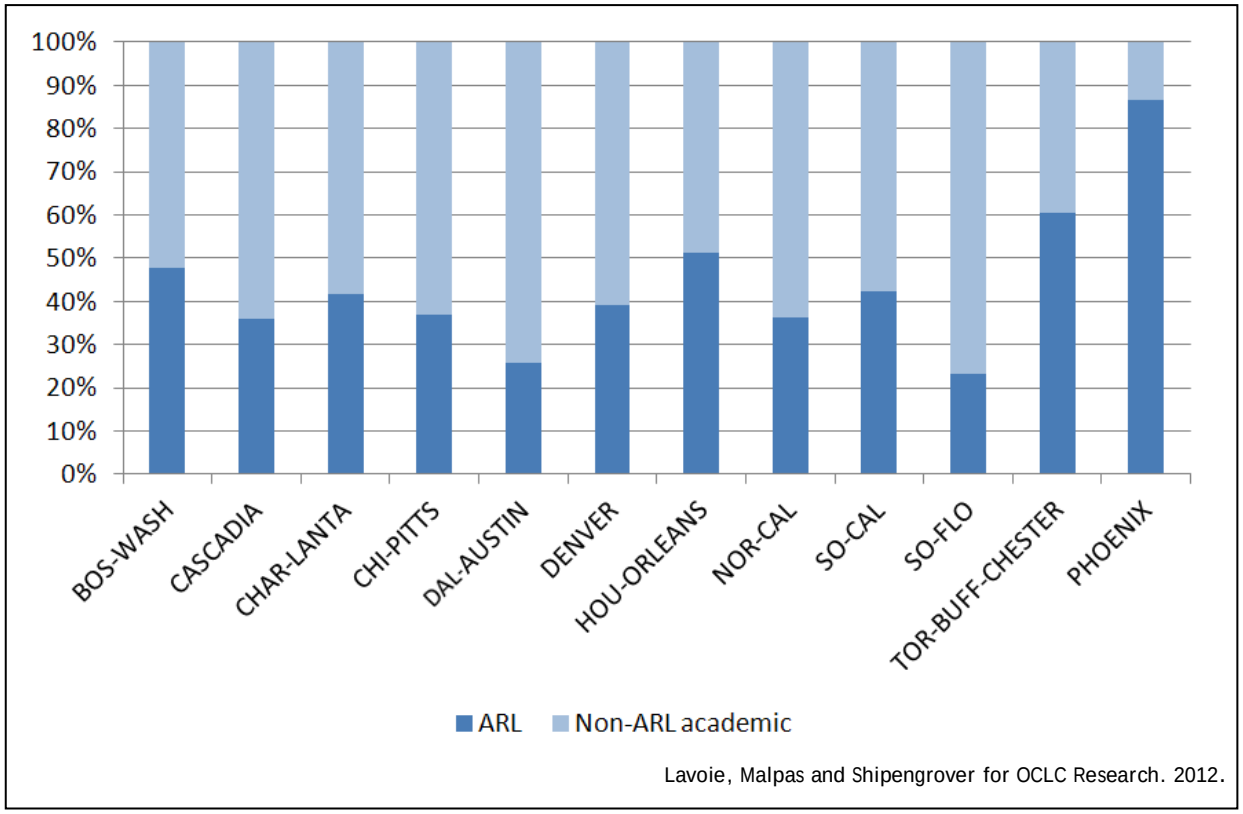


Figure 7. Share of ARLs in academic print book holdings, by region

The fact that most regional print book inventory is managed by academic institutions suggests that regional print book collections are, on average, geared toward the needs of faculty and students in higher education. This is further evidenced by the relatively low percentage of print book holdings belonging to public libraries in most regions (see figure 6): in half the regions, the share of public libraries is below a quarter, and in several regions the share is particularly low (BOS-WASH and TOR-BUFF-CHESTER, both at 17 percent). However, a few regions do exhibit relatively high percentages of public library print book holdings: SO-FLO (43 percent); CHI-PITTS (36 percent); and PHOENIX (35 percent). These regional collections would seem to be better positioned, vis-à-vis other regions, to serve the needs of general users.

Rareness is common within and across regional collections

WorldCat holdings data suggests that a significant share of print book inventory is relatively scarce both within regions and across regions. At least three quarters of the print book publications in each regional collection can be found at five or fewer institutions in the

region. Recall that a print book publication is a distinct imprint or edition of a printed book. Therefore, other publications pertaining to the same work may be available at other institutions. For example, while a particular publication of *A Tale of Two Cities* may be rare in the sense that it is held by only three institutions in the BOS-WASH region, many other publications of the same work may be available at other institutions in the region. Moreover, a print book holding indicates that an institution holds at least one copy of the publication in question; it may be that the institution holds many copies, which would alleviate to some degree the apparent scarcity observed at the publication level.

In some regions, the percentage of print book publications held by five or fewer institutions is particularly high: in DENVER, it reaches 89 percent, while in PHOENIX it is 95 percent. A partial explanation for exceptionally high percentages of "rare"[15] publications (that is, held by five or fewer institutions) might be found in a correspondingly high fraction of print book holdings within the region associated with ARL institutions. Intuition would suggest that the largest research libraries are likely to possess relatively unique print book collections vis-à-vis other institutions. In fact, the percentage of rare publications in a region and the share of print book holdings belonging to ARL institutions do exhibit a moderate degree of positive correlation,[16] indicating that regions with a relatively heavy ARL presence tend to have higher shares of rare materials.

The apparent "lack of abundance" of many print book publications within regional collections suggests both opportunity and challenges. Low levels of duplication correspond to high levels of uniqueness within a regional collection, which in turn suggests that a regionally consolidated collection would represent a significantly richer information resource, in terms of scope and depth, than what is available at any single institution. However, the ability to capitalize on this uniqueness—and confer benefits on regional users—will depend on the geographic size of the region and the robustness of its inter-lending infrastructure. Potential benefits will also be scaled to the extent that aggregate regional demand for a particular print book publication exceeds local demand at the institution or institutions where the publication is held.

Rareness is also common *across* regional collections. Forty-nine percent of the publications in the North American print book collection are available only in one regional collection, or are available only in either the US or Canadian "extra-regional" collection.[17] Eighty percent of the publications are available only in five or fewer regions.[18] Significant portions of several regional collections are unique to their regions: a third of the BOS-WASH collection, and a quarter of the TOR-BUFF-CHESTER collection, can be found in no other region. The majority of the regionally unique materials are concentrated in regions located in the eastern half of the United States and Canada; more specifically, about 70 percent of the regionally unique materials are located east of the Mississippi River.

Scarcity or uniqueness within a region does not seem to be a predictor of scarcity or uniqueness across regions. As figure 8 shows, a strong relationship between these characteristics is not apparent. In fact, if any relationship exists at all, it appears to be a negative one: regions with higher levels of intra-regional uniqueness tend to have relatively fewer materials unique to the region. This counter-intuitive relationship seems to be driven by regional size. Regions located to the upper left on the chart tend to be smaller: PHOENIX, DENVER, DAL-AUSTIN, CASCADIA; regions located toward the lower right tend to be larger: BOS-WASH, TOR-BUFF-CHESTER, CHI-PITTS. A possible explanation for the pattern in figure 8 is that smaller regions tend to have fewer institutions, which may act to reduce rates of duplication within the region. On the other hand, fewer institutions also means fewer materials in the regional collection, and therefore fewer opportunities to include rare or unique publications not available in other regions.

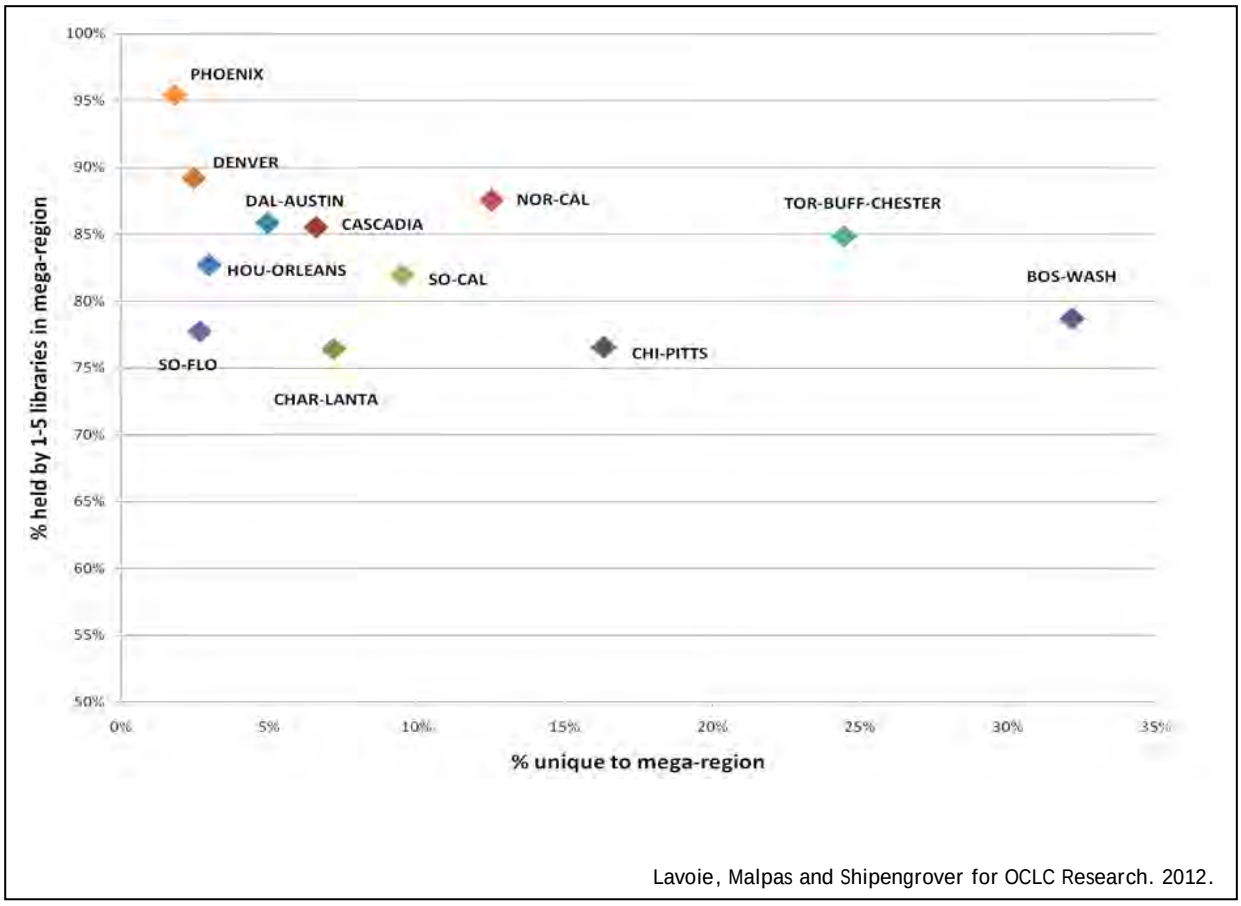


Figure 8. "Rareness" at the intra-region and inter-region levels

Analysis of overlap within and across regions indicates that considerable distinctiveness attaches to the regional collections at several levels. Consolidation at the regional level yields

an aggregate print book resource that is richer in scope and depth than any single local collection. But distinctiveness also manifests at the inter-regional level, where a significant portion of the overall North American print book collection is available in only a few or even a single region. It is worth noting that no regional collection is completely subsumed within another regional collection, or can be entirely duplicated through the combined holdings of a group of regions. All regional collections have a store of print book publications that are unique to that region. Even the smallest regional collection—PHOENIX—contains a fraction of materials (2 percent, or nearly 70,000 distinct print book publications) that are only available in that region.

Regional collections are globally diverse and exhibit similar collecting patterns across broad subject areas

Global diversity is a characteristic common across all regional collections, measured by the presence of non-English language materials and books published outside North America. Each of the twelve mega-regional print book collections included well over 200 countries of publication, with the highest total (247) found in BOS-WASH. Similarly, the publications in each regional collection reflected a wide range of languages, although the cross-regional variation in the number of languages was higher than for countries of publication. The region with the most languages represented in its collection was BOS-WASH with 473; the regional collection with the fewest number of languages, DENVER, had no less than 265 languages represented. Figure 9 shows the percentages of each regional print book collection published outside North America, and published in a language other than English. As the figure illustrates, non-North American and non-English print book publications account for significant portions of each regional collection.

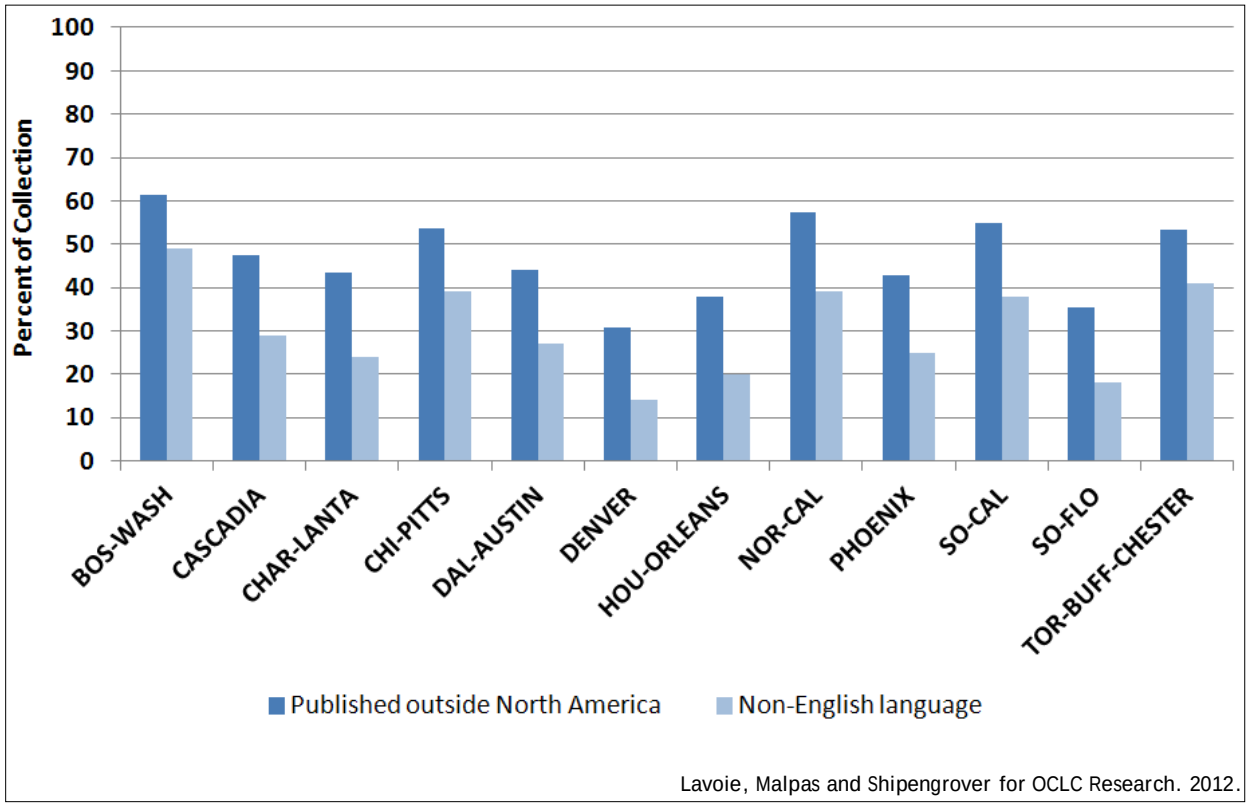


Figure 9. Global diversity in regional collections

The mega-regional print book collections display considerable similarity in regard to their subject make-up. A detailed subject analysis of the twelve collections could fill another paper; here, we limit ourselves to a few rough but indicative measures, based on analyses using the OCLC Conspectus[19] and FAST[20] topical subject headings in WorldCat records.

All of the regional collections had "Language, Linguistics, and Literature" as the most frequently assigned Conspectus division, and "History and Auxiliary Sciences" as the second most frequently occurring division. Ten of the twelve regions had "Philosophy and Religion" as the third most frequently occurring division. "Business and Economics" and either "Art and Architecture" or "Engineering and Technology" generally rounded out the top five Conspectus headings for the regional collections. The top five most frequently occurring divisions accounted for anywhere from 45 percent (BOS-WASH) to 60 percent (DAL-AUSTIN) of print books with Conspectus headings assigned. With the exception of BOS-WASH, the top five divisions accounted for more than half of the publications in the regional collections, suggesting a "long tail" shape to the distribution across Conspectus divisions in each region.

Similar results are obtained through an analysis of FAST topical headings, which characterize subjects at a more granular level. In each region, "political science" was the most frequently

encountered topical heading, with little cross-regional variation in the top five topical headings. Expanding to the top 250 topical subject headings in each region provides further evidence of cross-regional similarity in subject representation. As table 4 shows, the degree of intersection between each region's top 250 topical subject headings, and a benchmark top 250 list for the North American print book collection as a whole is substantial. No region overlaps less than about three-quarters with the benchmark collection, and for most regions, the overlap is much greater. In short, no region appears as an "outlier" in terms of an exceptionally distinctive subject composition for its regional print book collection.

Table 4. Regional overlap of top 250 most frequently occurring topical subject headings with North American print book collection[21]

Region	Intersection	Overlap (%)
BOS-WASH	228	91
CASCADIA	203	81
CHAR-LANTA	206	82
CHI-PITTS	225	90
DAL-AUSTIN	199	80
DENVER	184	74
HOU-ORLEANS	192	77
NOR-CAL	211	84
PHOENIX	187	75
SO-CAL	204	82
SO-FLO	189	76
TOR-BUFF-CHESTER	212	85

One distinction that is discernable across regional collections in terms of subject composition is that regional collections tend to have a regional flavor. FAST subject facet distributions show that regional collections collect relatively heavily in subject areas related to the region itself, pertaining to geography, local history, local events, and so on. For example, in DAL-AUSTIN, books about Texas and the Mexican War (1846–1848) figure prominently relative to other regions; similarly, in SO-FLO, relatively heavy concentrations of books can be found about Florida and Cuba. So while all regions seem to collect materials in the same general topical areas (perhaps reflecting production trends or other meta-regional trends), each seems to also specialize in books about region-specific subjects.

Taken together, this stylized fact, which identifies the subject-based similarity between the regional collections, and the previous one that asserts that "rareness is common" both within and across regional collections seem at first glance contradictory. Yet these stylized facts co-exist quite easily. A useful metaphor to illustrate this is to consider the regional collections as a group of retail stores that, by and large, sell the same type of merchandise, but tend to carry different brands of that merchandise. So while all the regional collections have relatively similar subject distributions, they also display a significant degree of distinctiveness in terms of individual offerings within subject areas—that is, specific publications. Although we did not examine subject data at the level of an individual institution's holdings, we would hypothesize that this phenomenon also holds within regions—i.e., similar subject distributions across institutions, but distinctive individual offerings within subject areas. This suggests that institutional collections consolidated at the regional level, or regional collections consolidated at a national or international level, result in a richer set of offerings within a shared pattern of collecting across broad subject areas.

Size is a driver for uniqueness, diversity, and age as characteristics of collections

The twelve regional collections vary considerably in size, with the largest, BOS-WASH, nearly seven times the size of the smallest, PHOENIX. Given that the larger regional collections by definition have more publications than smaller collections, it is easy to predict—and to confirm—that the larger collections will have greater numbers of publications that are unique to their region; greater numbers of publications originating from countries outside North America, or published in languages other than English; and greater numbers of older publications, than the smaller collections. Examination of the data shows, however, that in fact the difference between large and small regions on these points is more fundamental: differences can be detected not just in absolute terms, but in terms of proportions, which in turn suggests differences in collecting behavior.

The three largest regional collections—BOS-WASH, CHI-PITTS, and TOR-BUFF-CHESTER—unsurprisingly have the largest numbers of print book publications that are unique to their respective regions. By "unique" we mean that according to WorldCat holdings data, a particular print book publication is present in the collection of a single region. But these regions also exhibit the highest proportion of their collections corresponding to materials unique to the region: 32 percent for BOS-WASH; 16 percent for CHI-PITTS, and 24 percent for TOR-BUFF-CHESTER. On the other end of the spectrum, the three smallest regional collections—PHOENIX, DENVER, and SO-FLO—exhibit the three lowest proportions of regionally-unique materials: 2 percent, 2 percent, and 3 percent, respectively. This suggests that the presence of unique materials is proportionately less in smaller collections, or in other words, such materials are collected less intensively than in larger regions.

Similar results pertain to the presence of print books published outside North America, and those published in languages other than English. Again, the three largest regional collections exhibit high proportions of these materials, with BOS-WASH devoting 61 percent and 49 percent to non-North American and non-English materials, respectively; CHI-PITTS, 54 percent and 39 percent; and TOR-BUFF-CHESTER, 53 percent and 41 percent. The smallest collections, on the other hand, displayed relatively small proportions of these materials: PHOENIX, 43 percent and 25 percent; DENVER, 31 percent and 14 percent; and SO-FLO, 35 percent and 18 percent.

Figure 10 summarizes the percentages of each regional collection corresponding to publications that are regionally unique; published outside North America; and published in a language other than English. The regional collections are ordered from largest to smallest. The figure demonstrates the strong correlation between size of the collection, and its propensity to be comparatively unique and globally diverse. There are some exceptions, of course—for example, the PHOENIX regional collection seems to be more globally diverse than its size would predict[22]—but in general, larger collections have greater shares devoted to regionally-unique and globally diverse publications that smaller ones.

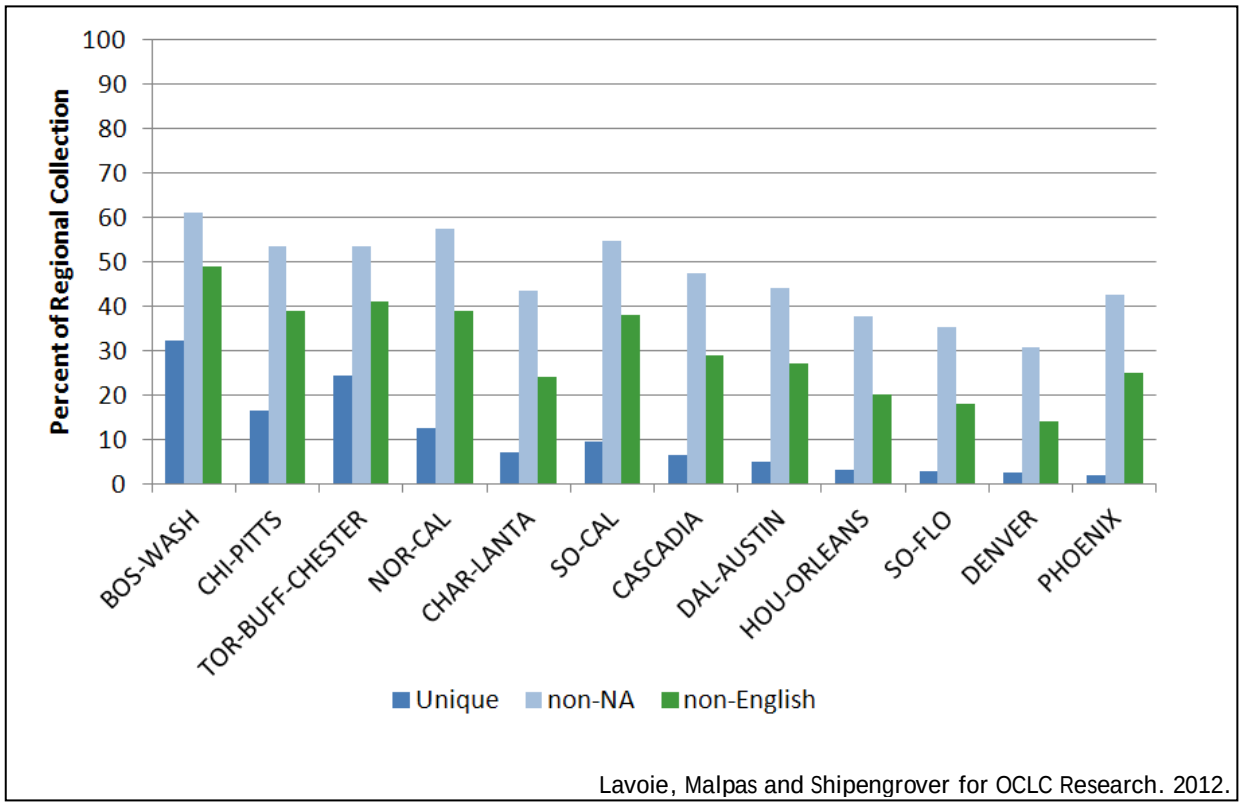


Figure 10. Uniqueness and global diversity as percentages of regional collections

Median age of the print book publications in regional collections is also strongly correlated with size. BOS-WASH, the largest collection, also has the "oldest" collection with a median

age (years since publication) of 34 years. CHI-PITTS, the second-largest collection, also has the second-oldest with a median age of 30 years. The "youngest" collections, on the other hand, are the smallest: PHOENIX at 23 years; SO-FLO at 24 years; and DENVER at 25 years.[23]

In summary, our findings suggest that large regions not only collect larger numbers of unique and globally diverse materials, but in general they collect proportionally more of these materials than small regions. Moreover, the higher median age of the larger regional collections suggests they tend to either collect proportionally more older materials, or retain materials in their collections for longer periods—or both. We conclude, therefore, that regional collection size seems to be a driver for uniqueness, global diversity, and age.[24]

The Largest Regional Collections Can Serve as Rough Substitutes for Smaller Regional Collections

Although each regional collection has a distinctive contribution to make to a North American network of consolidated regional collections, a pair-wise overlap comparison across regional collections reveals that the largest regional collections subsume most—although not all—of the print book publications available in the smaller regional collections. In this sense, the largest collections closely approximate, and therefore could serve as reasonable substitutes for, the smaller collections.

Bilateral overlap comparisons across the twelve regions reveal some interesting patterns. The BOS-WASH regional collection stood out as the collection subsuming the highest portions of the other regional collections. As figure 11 illustrates, BOS-WASH subsumed at least three quarters of nine of the other eleven regional collections. In addition, BOS-WASH accounts for 70 percent of the CHI-PITTS collection, and 65 percent of the TOR-BUFF-CHESTER collection. Some of the smaller regional collections are almost entirely subsumed within the BOS-WASH collection. For example, 95 percent of the PHOENIX collection, 93 percent of the DENVER collection, 92 percent of the HOU-ORLEANS collection, and 92 percent of the SO-FLO collection is duplicated within the BOS-WASH collection.

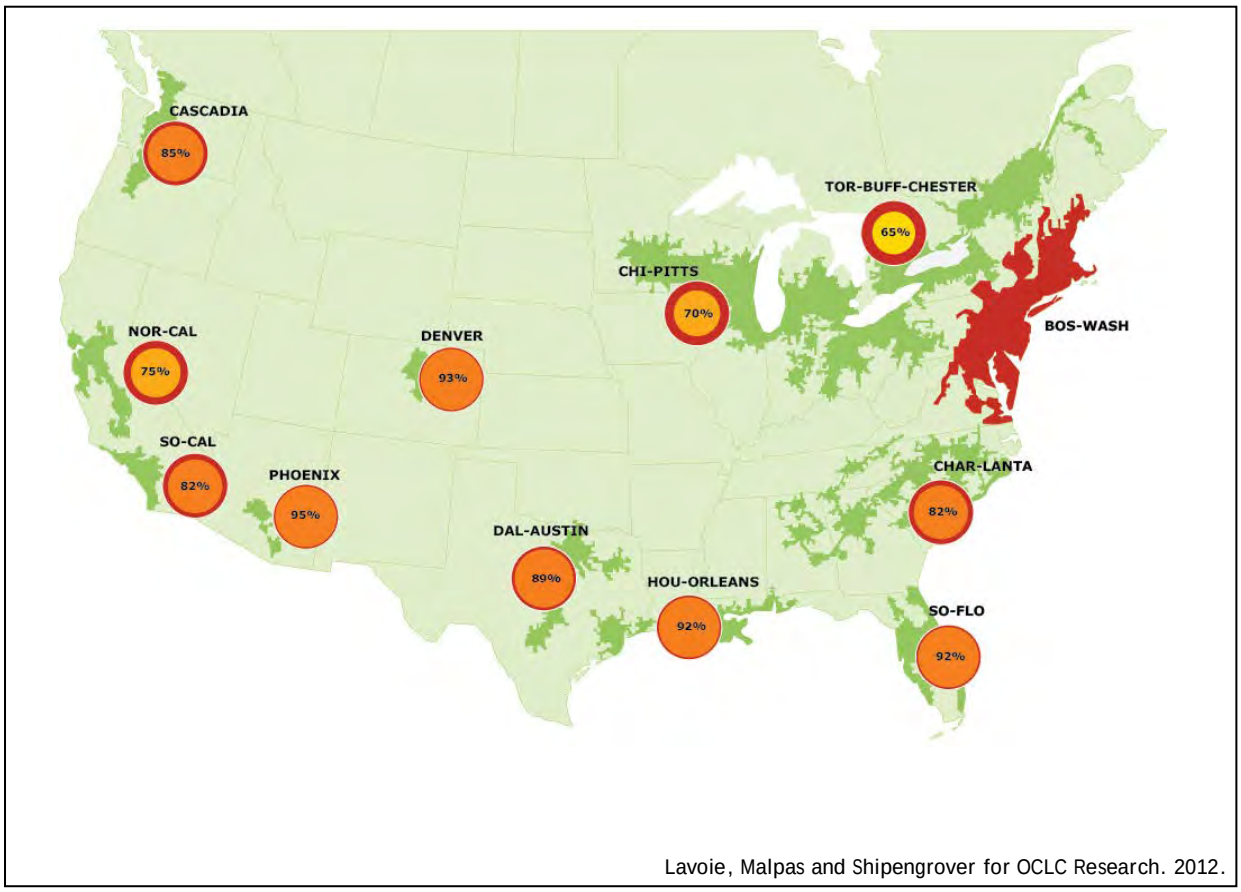
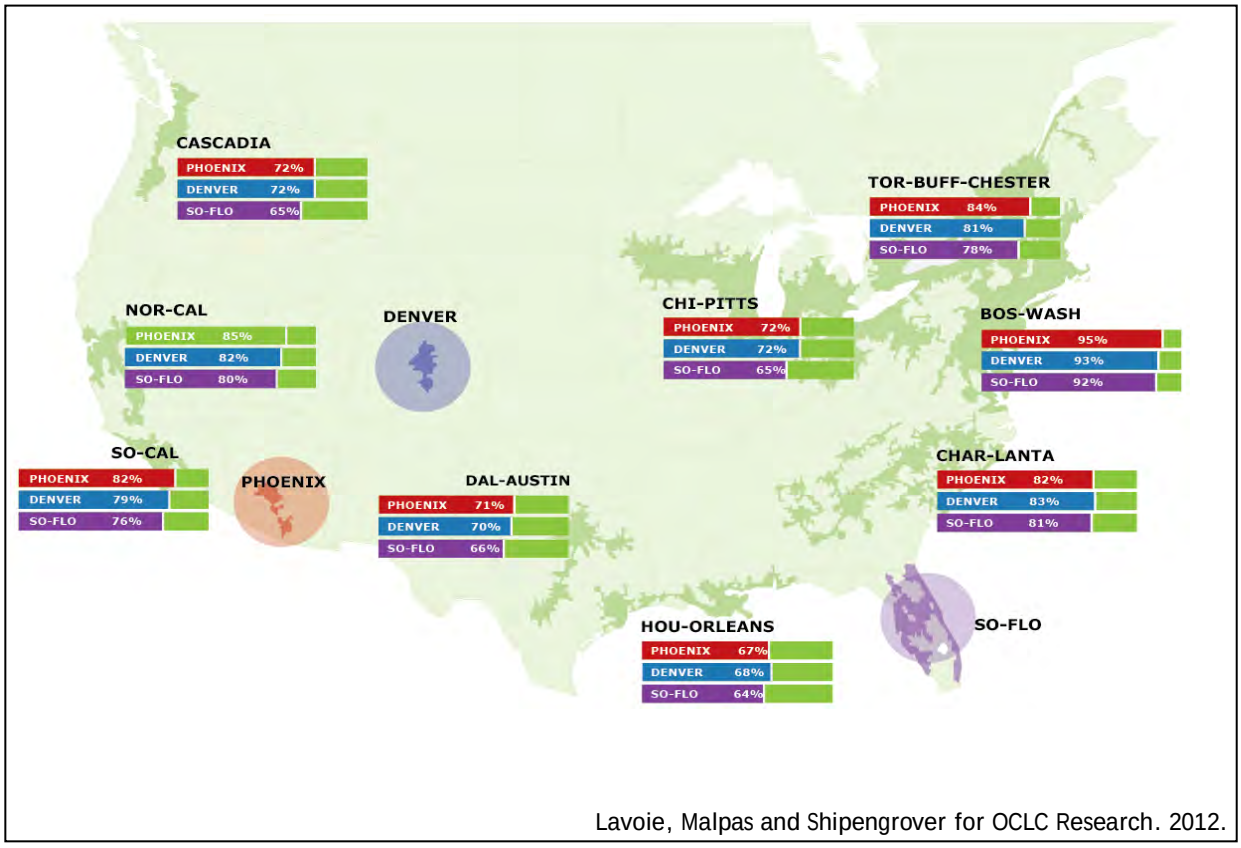


Figure 11. Bi-lateral overlap with the BOS-WASH collection, by region

More generally, we find that as a rule, the smallest regional collections are roughly duplicated by the largest collections. Figure 11 illustrates this in the context of the three smallest regional collections: PHOENIX, DENVER, and SO-FLO. Eighty percent or more of the PHOENIX collection is duplicated within six other regional collections; 80 percent or more of the DENVER collection is duplicated within five other regional collections (and 93 percent within BOS-WASH alone); 80 percent or more of the SO-FLO collection is duplicated within four other regional collections. Significantly, each of the three small regions is geographically near to at least one other region that overlaps with at least 80 percent of its collection. This would likely ease the logistical challenges involved with one of the small regions partnering with a larger neighbor to meet some part of its print book management and access needs.



Lavoie, Malpas and Shipengrover for OCLC Research. 2012.

Figure 12. PHOENIX, DENVER, and SO-FLO overlap with other regional collections

An important take-away from this analysis is that all of the regional collections can be paired with another collection that overlaps significantly with the first collection's holdings, and more specifically, the BOS-WASH collection is the highest overlapping collection for all the other regional collections (see figure 11)[25]. But it is equally important to note that no regional collection is completely subsumed within a larger collection. Each collection includes a slice that is found nowhere else, and therefore represents a unique contribution to the overall North American print book collection.

Another way of considering cross-region overlap is to examine the degree to which the cumulative holdings of the regional collections progressively cover the extent of the overall North American print book collection of 45.7 million publications. Table 5 presents the cumulative coverage of the North American collection, by the twelve regional collections, ranked from largest to smallest.

Table 5. Cumulative coverage of the North American print book collection

Region	Print Book Publications	Percent of N. American	Percent increase
BOS-WASH	26,105,425	0.57	--
CHI-PITTS	31,699,504	0.69	0.12
TOR-BUFF-CHESTER	36,246,732	0.79	0.10
NOR-CAL	38,303,267	0.84	0.05
CHAR-LANTA	39,223,404	0.86	0.02
SO-CAL	40,310,654	0.88	0.02
CASCADIA	40,851,328	0.89	0.01
DAL-AUSTIN	41,245,684	0.90	0.01
HOU-ORLEANS	41,433,933	0.91	< 0.01
SO-FLO	41,598,163	0.91	< 0.01
DENVER	41,719,110	0.91	< 0.01
PHOENIX	41,800,348	0.92	< 0.01

The findings reported in table 5 suggest a highly skewed distribution of the North American print book resource across the twelve mega-regions, with the vast majority of the North American collection concentrated in a few of the largest regions, and these large regions concentrated in the eastern half of the US and Canada. The BOS-WASH collection, containing 26.1 million print book publications, alone comprises 57 percent of the overall North American print book collection. Adding the next largest collection, CHI-PITTS, increases coverage of the North American collection to 69 percent. The three largest regional collections together account for almost 80 percent of the North American collection, while the six largest collections cover nearly 90 percent. It is clear from table 5 that diminishing returns set in quickly as new collections are added to the cumulative total. Taken together, the twelve mega-regional collections account for 92 percent of the overall North American print book collection; the remaining 8 percent is located in the US and Canadian areas outside the mega-regions.

Digital surrogates exist for significant portions of the regional collections

Strategies chosen for managing legacy print collections will depend in part on the availability of digital surrogates for print book publications. Digitization of print books has been preceding apace for several years now, under the auspices of mass digitization programs such as Google Books. As the volume of digitized materials has increased,

services such as HathiTrust have emerged to manage and share this content. Reliable access to digital surrogates creates a strong incentive to reduce the scale of print book inventory managed locally.

HathiTrust is a digital archiving service that manages digitized materials on behalf of more than sixty partner institutions (primarily American universities). As of March 2012, HathiTrust reports its collection to include about 5.4 million digitized book titles.[26] Table 6 reports HathiTrust's coverage of each of the twelve regional print book collections.

Table 6. HathiTrust coverage of regional print book collections

Region	Publications	In HathiTrust	Share of Regional Collection
BOS-WASH	26,105,425	3,719,184	0.14
CASCADIA	6,987,064	1,977,901	0.28
CHAR-LANTA	10,156,810	2,240,706	0.22
CHI-PITTS	18,558,201	3,700,089	0.20
DAL-AUSTIN	6,383,756	1,790,966	0.28
DENVER	4,047,196	1,216,497	0.30
HOU-ORLEANS	5,162,621	1,470,582	0.28
NOR-CAL	12,481,999	2,927,296	0.23
PHOENIX	3,827,173	1,258,297	0.33
SO-CAL	9,771,974	2,433,638	0.25
SO-FL	5,008,657	1,286,838	0.26
TOR-BUFF-CHESTER	14,699,921	2,827,021	0.19

The results in table 6 indicate that the HathiTrust corpus of digitized books accounts for significant portions of the regional collections. The PHOENIX and DENVER regions stand out in this regard, with about a third of their respective collections represented in HathiTrust. Seven of the twelve regions have a quarter or more of their collections in HathiTrust, while all except two regions have at least 20 percent. The outlier is BOS-WASH at 14 percent, although this result is primarily due to the size of the BOS-WASH collection; in absolute terms, BOS-WASH has the largest number of publications with digital surrogates in HathiTrust.

Examination of the subject content of the portion of the HathiTrust collection overlapping with the North American print book collection suggests a weighting toward the humanities. The three Conspectus subject areas "Language, Linguistics and Literature," "History and Auxiliary Sciences," and "Philosophy and Religion" account for half of the publications in

HathiTrust overlapping with a print publication in the North American collection; in contrast, these subject areas account for 32 percent of the North American collection itself. Several factors may account for this. One has to do with collecting patterns on the part of HathiTrust contributors. Most of the Hathi collection was contributed by large research universities, and there is some evidence that institutions of this kind tend to have these subject areas more prominently represented in their collections (Lavoie and Dempsey 2009). Second, there are indications that some of the contributors to the HathiTrust collection made a special effort to provide materials that were likely in the public domain. By definition, these would be older materials, many pre-dating 1923. It is possible that institutions are more likely to retain older humanities-related materials—e.g., literature, works of history, etc.—than older materials in the sciences, which have a faster rate of obsolescence.

Future strategies for managing print book collections will hinge on the library community's capacity to provide access to alternative formats, including digital surrogates. Digitized texts offer a broad range of features and conveniences to readers in comparison to the print originals, and could support a transformation of library operating models, enabling broader access to the collective resource while also reducing costs associated with managing redundant physical inventory. The current state of HathiTrust coverage of the twelve mega-regional print book collections, and the North American print book collection as a whole, suggests that significant progress has been made in this regard.

Print book publications held outside the mega-regions in the US and Canada constitute significant collections in their own right, although diffused over more institutions and a larger geographical space.

The analysis in this report focuses on the consolidated print book collections of the twelve North American mega-regions. As noted above, 92 percent of the overall North American print book collection is represented within these twelve collections. The remaining eight percent is found exclusively in US and Canadian print book holdings scattered across the space between the regions. While 8 percent may seem a small proportion, in absolute terms it represents nearly 4 million print book publication that are not available in any of the twelve regional collections. Moreover, the general characteristics of the materials in the US and Canadian "extra-regional" collections resemble those of a large and small regional collection, respectively. However, both of the extra-regional collections also have a few unique characteristics.

The US extra-regional collection consists of 15.7 million print book publications (distinct imprints or editions of books in printed form) which in comparison to the twelve mega-region collections, would make it the third largest collection after BOS-WASH and CHI-PITTS. The US extra-regional collection alone can account for about a third of the overall North American

collection. The more than 217 million total holdings in the US extra-regional collection—the sum total of the number of print book publications held in each institutional collection in the US extra-regional space—exceeds the total holdings of every regional collection; BOS-WASH has the next highest total at 191.6 million. As a consequence, the ratio of holdings to print book publications in the US extra-regional collection (13.83) is more than half again as high as the regional collection with the highest ratio (CHI-PITTS at 8.94). This in turn suggests a relative abundance of print book inventory (and higher level of duplication) in the area between the regions in the US.

The US extra-regional collection resembles the larger regional collections in terms of several key characteristics. Fifty percent of the collection was published outside North America, and 33 percent consists of non-English language materials, which would place the collection among the upper half of the regional collections with respect to both characteristics. The median age of the collection is 31 years, which makes it the second oldest collection after BOS-WASH (34 years). Fourteen percent of the collection is unique to the US extra-regional area.

The Canadian extra-regional collection encompasses 5.8 million publications and 14.8 million holdings, which places it on a scale similar to the smaller mega-region collections. The Canadian collection alone would cover about 13 percent of the overall North American print book collection. The percentages of materials published outside North America (43 percent) and in languages other than English (25 percent) are more in line with the smaller mega-region collections than the larger ones. But the Canadian collection does depart from the smaller regional collections in terms of its median age, which is 30 years, placing it among the "older" regional collections.

Based on available evidence, print book inventory appears to be relatively scarce in the Canadian extra-regional collection, with 89 percent of the publications available at five or fewer libraries, the second highest percentage after the PHOENIX region. The Canadian region differs from smaller mega-region collections in that a relatively high percentage of its collection is unique to the area (15 percent). This percentage would place it fourth among the mega-region collections, after the largest regions BOS-WASH, TOR-BUFF-CHESTER, and CHI-PITTS, and slightly ahead of the much-larger US extra-regional collection.

The extra-regional collections in the US and Canada constitute important collections in their own right, representing rich collections of print book publications with significant uniqueness vis-à-vis other regional collections. However, these collections are dispersed over a considerable geographic area with no natural framework within which to consolidate them into a single virtual collection. Within the vast extra-regional space, a number of successful and sometimes overlapping inter-lending networks do exist. It remains to be seen

if such networks could be effectively federated into a system that would leverage the highly diffused aggregate print book resource as a regional asset. In short, the extra-regional collections are a varied and abundant inventory that is not organized in a way that easily supports shared access or management. The total value of this resource is imperfectly reflected in current management and inter-lending systems; it is in effect an "unaddressable resource" in the network.

A number of significant concentrations of print books can be found outside the mega-regions. Figure 12 indicates the locations of the top five largest concentrations of print book holdings outside the mega-regions in both the US and Canada. In the US, the comparison was conducted on the basis of geographical areas associated with US Postal Service sectional center facilities (identified by the first three digits of a US ZIP code)(Wikipedia 2012a). In Canada, the comparison was based on a similar geographical unit, the forward sortation area, defined by the first three characters of a Canadian postal code (Wikipedia 2012). Although much smaller than mega-regions, these geographical units were the largest identifiable areas that are distinct from—i.e., do not overlap with—the mega-regions.

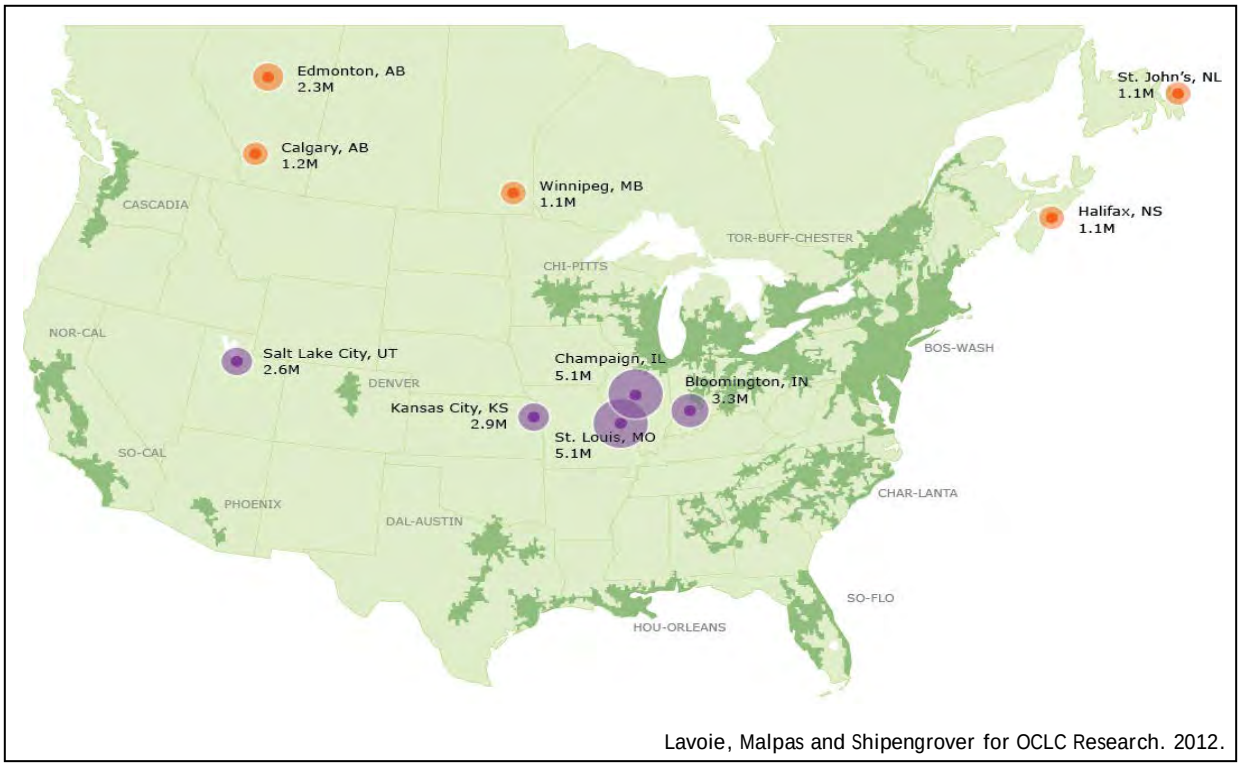


Figure 13. Top five concentrations of print book holdings outside the mega-regions, US and Canada

As figure 13 illustrates, the largest concentration of books outside the mega-regions is the St. Louis (Main) area, with more than 5 million print book holdings. Washington University and the

St. Louis Public Library have significant holdings in this area, as well as St. Louis University and the University of Missouri, St. Louis. Champaign (North) in Illinois is not far behind: the University of Illinois, as well as three public libraries—Urbana Free Library, Champaign Public Library, and the Danville Public Library—all have significant print book holdings in this area. In Canada, the largest concentration of print books outside the mega-regions is found in Edmonton (North Capilano), with nearly 2.3 million print book holdings—the University of Alberta is located there. Calgary (Kensington/Westmont/Parkdale/University) is the home of the second largest concentration in Canada with nearly 1.2 million holdings; the University of Calgary is the key institution in the area.

It is interesting to note that in the US, the second- and third-largest extra-regional concentrations of print book holdings—Champaign, Illinois and Bloomington, Indiana—are situated very close to the edge of a mega-region (CHI-PITTS). It is easy to imagine that institutions in these locations would find opportunities to connect themselves to cooperative print management efforts within CHI-PITTS. Indeed, the geographical distribution of the membership of the Committee on Institutional Cooperation (CIC), a collaborative body that includes the twelve Big Ten Conference institutions plus the University of Chicago, tracks very closely to the CHI-PITTS region. But the CIC also includes the University of Illinois and Indiana University, owners of the largest print collections in the Champaign and Bloomington areas, respectively. Moreover, CIC also includes more distant extra-regional institutions such as the University of Iowa and the University of Nebraska. Cooperative print management strategies within the CIC would therefore include institutions from outside CHI-PITTS, even though most members are located within the region or on its borders. In general, cooperative print management initiatives based on existing cooperative structures like CIC are likely to draw in partners from outside the mega-regions.

In all of the extra-regional concentrations of print book holdings shown in figure 13, an academic institution represents the entity in the area with the most print book holdings. This aligns with the stylized fact above that identifies academic institutions as the custodians of the majority of the print book inventory in the twelve regional collections. However, in all five of the US extra-regional areas shown in figure 13, a public library represents the second largest collection of print books in the area. In some of the areas, the share of holdings of this public library is significant: for example, 18 percent (Salt Lake City Public Library) and 15 percent (St. Louis Public Library). These results serve to remind us that public libraries are also key stakeholders in the future of print book collections. Regionally based cooperative print management strategies should not forget the potential contributions of public libraries to print consolidation activities.

Key Implications

A number of implications for the future of print management emerge from the stylized facts described above.

Stewardship of the North American print book collection will require coordination on a supra-institutional scale.

Our study reveals that the geographic distribution of the aggregate North American print book collection is remarkably uneven, with the largest regional collections—all located in the eastern half of the US and Canada—accounting for the vast majority of print book publications and holdings in the North American collection. At the same time, we discovered that scarcity of holdings on a regional basis is not a reliable indicator of scarcity elsewhere in the larger library system. Taken together, these facts suggest that effective stewardship of the aggregate resource will require both a supra-institutional view of the system-wide collection and cooperative agreements that transcend organizational and even geographic boundaries.

Individual institutions, and even regional cooperatives, are likely to misjudge the relative preservation risks to which their collections are exposed if they fail to appreciate the highly diffuse distribution of the collective print book collection. Just as supply chain management in the retail and manufacturing sectors is informed by a global view of demand, local and regional library inventory must be considered in the larger context of system-wide holdings and aggregate demand. Without this broader perspective, individual institutions run the risk of over-investing in local print preservation strategies at the expense of other, mission-critical services.

The appropriate scale of consolidation and cooperative management for regional print book collections will vary

There are many examples of collaborative print management schemes organized by consortia operating at the state and provincial level, from OhioLINK's long-running effort to coordinate the purchase of books-not-bought-in-Ohio (Gammon and Zeoli 2003), to more recent state- and province-wide initiatives to identify and preserve "last copies" of books in Illinois, New Jersey and Ontario.[27] Libraries in Florida and Maine are currently working to develop state-wide strategies and shared infrastructure to manage local print and journal collections.[28] These initiatives will produce real benefits for libraries operating within their respective geographic areas; yet, as the mega-regions analysis suggests, the boundaries of the system-wide library print collection do not necessarily align with state borders. To the extent that mega-regions represent a natural unit of economic

organization, undergirded by common social, cultural, and economic interests, one might expect cooperative print management solutions to operate at a similar scale.

Print books acquired by libraries in Northern California are, in the aggregate, different from print books acquired by libraries in Southern California; library collections in the western area of New York State will more closely resemble collections in Ontario than collections in Ithaca or Syracuse. Titles judged to be last copies on a state-wide scale may be relatively abundant if evaluated against regional or supra-regional holdings. Management schemes focused on rationalizing collections on a per-state level risk disrupting supply and demand patterns operating on the higher, mega-region scale. A holistic approach to managing the collective resource will require cooperation across multiple states. This presents obvious challenges, as the social and technical infrastructure that is needed to support coordinated management of library resources at this scale is relatively limited. We consider this issue in further detail in the concluding section.

Paradoxically, the fact that "rareness is common" in the North American print book collection may constitute the single greatest imperative to collective action. While individual libraries are generally pleased to know that some part of their collection is distinctive, rare, or even unique, the capacity of regions to secure the long-term preservation of all such resources is limited. For regions with relatively few natural preservation partners—institutions with a mandate or other motivation to assume curatorial responsibility on behalf of the collectivity—stewardship may prove a burdensome challenge. The relatively scarcity of holdings (multiple copies of discrete titles) on a regional basis suggests that efforts to preserve distinctive resources will need to scale up to include multiple regions, if only to distribute the total cost of the curatorial enterprise.

As noted above, 75 percent or more of the print books in any of the North American regions are held by 5 or fewer libraries within that region. The sheer volume of material that would appear to warrant special preservation investment would overwhelm library budgets in any one region. Regional library cooperatives would arguably do better to concentrate their limited resources on the preservation of print book publications that are scarce or unique in the system as a whole, and redistribute the cost of preserving more ubiquitous material across a broader range of regional interests.

Significant bilateral duplication between mega-regional print book collections suggests opportunity for inter-regional cooperation in print management and fulfillment services.

While intra-regional supply of print books is generally limited, we discovered surprisingly high rates of pair-wise overlap between the largest mega-regional collections and those of smaller

mega-regions. An implication of this is that smaller regions may find it advantageous to externalize some print management operations to larger regional partners, who can achieve greater economies of scale in preservation and access services. This has important implications for inter-regional shared print agreements and, more generally, for the limiting factors for large-scale print management regimes. If a robust, inter-regional library logistics infrastructure is available, it is conceivable that regions with relatively high rates of duplication vis-à-vis other regions may source preservation and fulfillment services elsewhere. While it is highly improbable that any region would outsource management of distinctive resources, it is logical to imagine that as fulfillment models continue to diversify—with an increasing range of digital surrogate, print-on-demand, and inter-lending supply options available for a growing part of the retrospective print collection—strategies for sourcing commodity content will also change. Some efforts along these lines are evident with consortial and state- or province-wide licensing of digitized books in aggregations like Early English Books Online (EEBO). Content that was once difficult or even impossible to acquire in print, or regarded as a distinctive institutional asset, has become increasingly ubiquitous and as such lends itself to collective management.

A related implication is that mega-regions with a high rate of inter-regional duplication (PHOENIX and SO-FLO, for instance) may have greater options for sourcing cooperative print management solutions than regions with relatively low inter-region duplication. Larger mega-regions, which tend to have a lower inter-region duplication rate, are likely to be viewed as natural suppliers of print access and preservation, whether or not they actively seek to assume such a role. MINITEX, a resource sharing service funded by the Minnesota Office of Higher Education, offers a potential model for multistate preservation and access services. Through state contracts, MINITEX provides libraries in North and South Dakota with inter-lending and document supply services that leverage the vast collections of the University of Minnesota-Twin Cities. Such multistate resource sharing arrangements may have a downside for supplier institutions, however, as contractual obligations are likely to constrain their options for reconfiguring their own print book inventory.

A substantial part of each regional print collection has been digitized, creating opportunities for a virtual redistribution of system-wide assets; even so, major challenges for regional print management remain.

Our investigation revealed a sizeable overlap between regional print collections and the mass-digitized corpus in the HathiTrust Digital Library, ranging from a low of 14 percent in BOS-WASH to a high of 33 percent in PHOENIX. As with the bilateral duplication analysis, where we observed that the largest regional collections subsumed most of the smaller collections, we found that smaller regional collections were more likely to be substantially duplicated in HathiTrust. This is a notable finding, as most of the content contributors to the HathiTrust

corpus are located in the larger mega-regions and a majority of the contributed content was sourced from three zones: CHI-PITTS, NOR-CAL and SO-CAL.[29] It is remarkable, for example, that regions like PHOENIX and DENVER, which have (as yet) contributed no content to the aggregation, have a higher than average duplication rate with HathiTrust. An implication of this is that these smaller regions may stand to benefit more from the "replacement" value of HathiTrust than larger regions, assuming the digitized corpus is ultimately made available as a source of surrogate supply. This is a vexed issue at present, given uncertainties about the likely outcome of a legal dispute between rights-holders, Google, and HathiTrust.

The growing overlap between digitized books and retrospective print book collections may have another, equally transformative effect on the library system. As shown above, the supply profile of regional collections varies dramatically across North America, with a relatively small number of regions concentrated in the Northeast, Upper Midwest and (to a lesser extent) West coast holding a significant share of the aggregate resource. The increasing "dematerialization" of this content may ultimately diminish the striking asymmetries in supply that currently characterize the North American print book collection. This in turn may create opportunities for institutions in mega-regions with comparatively small regional print book collections, as well as those located outside of the mega-regions, as new fulfillment options emerge that level the playing field by reducing reliance on traditional print distribution networks.

As an increasing part of the source material that informs scientific research and inspires creative innovation moves online, the infrastructure that supports knowledge creation and information exchange will become less dependent on locally concentrated inventory and more reliant on systems that improve the flow of content, both in print and online. The degree to which libraries stand to gain or lose from this transition will depend in large measure on the success (or failure) of library fulfillment channels to compete with alternative supply chains. This is especially true for digitized print books, most of which are subject to copyright protections that prohibit electronic redistribution. Regional inter-lending networks might provide the backbone for a more robust library logistics network, enabling a greater number of libraries to benefit from a diminishing but still widely dispersed print book inventory. Looking forward, it is reasonable to imagine that regions that are successful in implementing "flow"-based strategies for print management that maximize the total value of the aggregate print book resource will have a competitive advantage over regions where management of print resources devolves to individual, differently equipped institutions.

While the vast and still growing digitized corpus is likely to have a major impact on print management strategies on both an institutional and regional basis, one should not lose sight of the fact that a large part of the system-wide print book collection has not yet been

digitized. As of March 2012, we calculate the total number of print book publications in the HathiTrust digital library to be approximately 4.9 million, representing only about 11 percent of the 45.7 million distinct print book publications in North America.[30] This estimate excludes titles digitized by Google and other agents that are not part of the HathiTrust collection, but nevertheless represents the core of the mass digitized resource that can be said to have a specific utility—long-term digital preservation—to institutions seeking to reduce local print preservation investments. Consequently, for the many millions of print book publications not replicated in digital archives like HathiTrust, the range of possible preservation strategies is relatively limited. Our application of the mega-regions framework suggests that institutions in the BOS-WASH, TOR-BUFF-CHESTER and CHI-PITTS regions—where the total print book collection is very large and the overlap with HathiTrust is lower than average—will have fewer preservations options available to them than institutions in PHOENIX, DENVER or HOU-ORLEANS. One can infer from this that large-scale, multiregion cooperative preservation strategies will be needed if the aggregate print resource is to be secured for the long term as a collective resource.

Changes in the organization of higher education will have a profound effect on academic library infrastructure and the disposition of print books in particular.

While the majority of the North American print book inventory is held in academic libraries, the distribution by type of academic library differs significantly from one mega-region to the next. This has important implications for the redistribution of institutional investment in print collections, as colleges and universities with a primary mission of teaching and learning are increasingly focused on cost-effective provisioning of course materials, especially electronic content. Investment and attention that was once directed toward building and maintaining local print collections now has other aims. As a consequence, stewardship responsibility for the aggregate print collection is increasingly concentrated on a relatively small—and unevenly distributed—population of research-intensive institutions.

Academic research libraries, which typically view stewardship of the scholarly record as central to their academic mission, are challenged to uphold a preservation mandate that encompasses an increasingly diverse range of information resources. Print preservation is just one among many stewardship responsibilities, and at even the largest North American research universities it must compete with other institutional priorities for scarce attention and resources. Looking ahead, we can anticipate that changes in print management strategies in academic libraries outside the ARL sector will have a decisive impact on libraries within the ARL sector. As an increasing number of academic institutions begin to reduce retrospective inventory and purchase fewer print books, a small minority of libraries with a mission-driven commitment to preservation will feel compelled to step into the breach to ensure continuing access to and the long-term survival of the print published record.

In mega-regions where a majority share of the print book stock is managed by non-ARL academic libraries, it may prove difficult to identify an entity with a "manifest destiny" to assume responsibility for the regional print resource. An upside of this topsy-turvy world, in which the largest and best resourced libraries have less control over their destinies than nimble, smaller institutions that have divested from legacy print, may be that ARL institutions can capitalize on the increased reliance on their collections to negotiate new business and mutual aid agreements that may help redistribute the costs of long-term preservation.

Resources located outside of the established mega-regions may prove difficult to mobilize as a collective asset.

As noted in one of the stylized facts, a resource equal in size to the third largest mega-regional print book collection is located in the portions of the United States outside the twelve North American mega-regions. A significant print book resource also exists in the Canadian extra-regional area. While significant in terms of size (library holdings) and scope (number of publications), these extra-regional collections are purely notional, in the sense that they are not underpinned by any existing collaborative arrangements, shared infrastructure, or other substantive bonds of mutual interest. Nor would one expect that they would be, given that the extra-regional areas stretch across the lengths of the US and Canada, and encompass collecting institutions of all descriptions.

Mobilizing the regional print book collections as a collective resource will be difficult; mobilizing the print book holdings scattered across the areas outside the regions will be even more challenging. This suggests that the extra-regional print book resource is at greater risk than print books undergirded by the comparatively strong infrastructure—both physical and collaborative—found in the mega-regions that can serve as a starting point for building regionally based cooperative print management strategies. Moreover, the print books distributed across the vast extra-regional area, especially those that are not duplicated in other mega-regions, will be difficult to mobilize for use beyond their local settings in the absence of more robust library logistics.

In the US, a partial solution to this problem might lie in the fact that the largest extra-regional concentrations of print book holdings are often not far away from a mega-region. As figure 13 illustrates, three of the top five extra-regional print book concentrations are quite close to the CHI-PITTS region; the fourth (Kansas City) is situated mid-way between DENVER and CHI-PITTS, while the fifth is near DENVER. One could imagine clusters of print book holdings in the areas surrounding the borders of the mega-regions being incorporated into the collaborative print arrangements of the mega-region itself, although the "thickness" of this border area—i.e., the distance from the mega-region at which effective integration is still feasible—is an open question. In the case of CHI-PITTS, the border area

encompasses several universities that participate in the Committee on Institutional Cooperation, which serves as the cooperative infrastructure for a large part of the regional research enterprise. Indeed, this regional infrastructure also serves a number of institutions located far outside the CHI-PITTS area, including the Universities of Iowa and Nebraska. Similarly, one can look to WEST, the regional print archiving effort, as evidence that institutions located in the extra-regional zone can leverage cooperative infrastructure that is anchored within an adjacent mega-region.

This solution works less well in Canada, where, as figure 13 illustrates, the largest clusters of print book holdings outside the mega-regions are more isolated. In these circumstances, another possibility might be the development of "mini-regions," perhaps taking the form of a collaboration between institutions located in a metropolitan area and its immediate hinterland. For example, Edmonton is distant from any existing mega-region, but might consider a cooperative print strategy that includes academic and public libraries in Edmonton and the surrounding area, perhaps even extending down to include Calgary, another metropolitan area outside of a mega-region. In this way, locations outside the mega-regions might replicate to a degree the benefits inherent in regional cooperation, albeit at a smaller scale. One can point to the cooperative print strategy of the Ontario Council of University Libraries (OCUL), which leverages the strengths of institutions located in the metropolitan areas of Southern Ontario, as an exemplar of this kind of sub-regional strategy.

Canada is distinctive in having a strong regional cooperative library infrastructure that spans vast geographic areas, sometimes including universities and always encompassing populations located far from metropolitan centers, often with limited Internet connectivity. These supra-provincial agglomerations will be challenged to establish cooperative print management strategies that can balance the interests of libraries hoping to reduce their print footprint and communities that have a reasonable expectation of equitable access to information. For example, the Consortium of Prairie and Pacific University Libraries (COPPUL) supports 22 academic institutions in four Western provinces, covering a landmass of nearly 75 thousand miles—a significant part of the extra-regional Canadian print book collection. COPPUL members have initiated a plan for shared management of a print journal archive, focusing on titles that are available in electronic format (COPPUL 2011). Extending this model to the monographic literature may prove difficult, given the geographic scale of the region and the comparatively attenuated infrastructure available to support a "flow"-based model.

Whether through alliance with a geographically proximate mega-region, the cultivation of a "mini-region," or some other strategy, the potential for considerable system-wide benefits is created by incorporating as large a share of the extra-regional print book resource as possible

into some form of multi-institutional collaborative arrangement for print management and access. Library consortia serving institutions located outside the mega-regions will have the opportunity to model a variety of potential solutions and best practices.

Optimizing print collections on a mega-regional basis will have system-wide impact.

The high concentration of print book inventory in BOS-WASH, CHI-PITTS and TOR-BUFF-CHESTER, and the relatively elevated levels of duplication in holdings across these regions, suggest that a large-scale cooperative effort to optimize and secure this resource against potential loss will deliver benefit to the library system as a whole. All libraries in North America benefit from the availability of these legacy collections, and their longevity will depend in part on the economic sustainability of the preservation and access that ensure their continued usefulness. Simply put, greater economies of scale can be achieved in regional approaches to print management that leverage existing infrastructure and inventory, than in approaches that require a massive redistribution of inventory.

Since there is a comparatively high level of duplication within and across these three mega-regions, and because most of the inventory is held by academic libraries that are, individually and collectively, revisiting their long-term investment in print collections, it will be necessary to guard against the uncoordinated withdrawal of materials upon which the North American library system as a whole depends. The consequences of a disorderly "draw down" of library holdings in BOS-WASH, CHI-PITTS and TOR-BUFF-CHESTER, which collectively hold 80 percent of the North American print book resource, would be felt across the library system as a whole. Thus, while significant system-wide benefit might be achieved through a deliberate rationalization of this supra-regional resource, significant harm may result if institutions continue to view print management (including the withdrawal of materials) as a purely local concern.

A recent survey of academic library directors in the US found that while about half of respondents felt that they lacked sufficient data to make informed judgments about the withdrawal of print journal back-files, more than 90 percent were planning or already actively engaged in projects to de-accession local serial holdings (Long and Schonfeld 2010). The same survey reported that, while academic library directors did not view a transition to reliance on e-books as imminent, a considerable majority (74 percent) "said that the withdrawal of print books would be an important strategy for their libraries in the future" if appropriate preservation and access mechanisms were in place (p. 36). The emergence of successful business ventures supporting the managed de-selection of print books in academic libraries suggests that this trend is already well underway, even where cooperative preservation infrastructure is still lacking.[31]

As libraries look to optimize the amount of local print inventory, it will be important to balance institutional imperatives to maximize library space recovery by reducing local physical holdings, with the core library mission of broadening access to information. Our investigation has revealed that each of the twelve mega-regions hold some distinctive print resources, assets that are not duplicated in any other regional collection and that add richness to the system-wide print book resource. Improving the "flow" or circulation of these resources—whether in print or digital form—will benefit the system as a whole by ensuring that the total value of library investment is effectively leveraged.

Looking at the aggregate print book collection from a supply-side perspective, it is clear that the relatively diffuse distribution of inventory is not optimized for fulfillment except perhaps at the local level. Inter-lending networks support a certain amount of load-leveling among potential suppliers, but they presently lack the system-intelligence or "shelf-awareness" that would enable a more dynamic alignment of supply and demand. As currently configured, library logistics are highly inefficient: the transaction costs of distributed print fulfillment are high (estimated at $30 per interlibrary loan) and relatively inelastic, in part because the sources of supply are not located where the demand originates.[32] It is not possible to aggregate potential users around existing sources of supply, so instead solutions must be found to improve the flow of resources to users. A flow-based approach to consolidation that leverages aggregate demand will amplify the impact of library investment in collections within and across mega-regions, while also enabling institutions located outside of these zones to benefit from greater economies of scale.

Conclusions

In this study we have applied the mega-regions framework to the distribution of the aggregate print book resource in North America, and considered some implications for cooperative approaches to print management. Among our key findings is that the variable distribution of print book publications (distinct imprints or editions of books in printed form) across mega-regions—with a few mega-regions accounting for a majority of the North American print book collective collection—along with variations in the characteristics of regional collections, is likely to result in divergent regional strategies for print management. At the same time, we find that the growing overlap between locally managed print book collections and collectively managed aggregations of digitized books may have a leveling effect, enabling regions with comparatively small collections to achieve greater efficiencies in print management than is feasible in the mega-regions with larger collections. We also find that the aggregate print book collection located outside of established mega-regions may prove difficult to leverage as a collective resource, except at a scale commensurate with existing social and technical infrastructure.

This report provides a supply-side picture of the North American print book collection, mapped against clusters of population and economic activity—i.e., the mega-regions. More work is needed to understand demand patterns within, across and outside of the North American mega-regions. It is hoped that regional consortia participating in intra-consortium borrowing programs—including Borrow Direct, OhioLINK p-circ, UC Request, and the Orbis-Cascade Alliance's Navigator system, among others—will begin to share and analyze inter-lending data to build a common understanding of aggregate demand patterns within and across regions. Equally important is additional evidence and analysis to improve our understanding of how collections located in the vast extra-regional areas function in the larger picture of supply and demand. Does the higher rate of duplication in collections in this zone correspond to a greater aggregate (and more diffuse) demand than that found in the mega-regions? Or does the additional inventory dispersed across this less populous area compensate for weaker distribution networks? A more complete picture of the aggregate demand characteristics within and outside of the mega-regions will help to address the striking asymmetries in supply that the present study has revealed.

An important corollary to the findings discussed in this report is that existing cooperative infrastructure may not be equal to the task of managing print resources at a mega-regional scale. The absence of a cooperative infrastructure that is fit-to-purpose for achieving an integrated regional print management strategy, or negotiating on behalf of regional partners, represents a significant constraint on the development of a system-wide, multiregional preservation plan. It remains to be seen if existing organizational structures will adapt to serve the growing need for supra-institutional and supra-regional planning and governance, or if new organizations will be needed to bridge the gap between state- or province-level and larger-scale federal or national approaches to cooperative print management. To the extent that regional consolidation of print resources enables individual institutions to reduce costly duplication in infrastructure and management, while maximizing the value of the collective resource, it seems likely—if not inevitable—that appropriate cooperative infrastructure will emerge, whether through a pragmatic process of boot-strapping or a top-down initiative to institutionalize shared print management.

While a few library consortia can reasonably claim to represent the interests of some institutions within the mega-regions—the Committee for Institutional Cooperation (CIC) maps reasonably well to CHI-PITTS, the Association for Southeastern Research Libraries (ASERL) aligns with a large part of CHAR-LANTA, the Orbis-Cascade Alliance is a close match to CASCADIA—and while each of these consortia is actively pursuing regional collection management initiatives, none has explicitly embraced a membership model or governance structure that would enable a supra-regional approach to managing the aggregate print resource. Moreover, none of these consortia include non-academic (e.g., public or corporate) libraries in their membership, so the nature of the aggregate collection they can marshal as a

cooperative resource is defined—and limited—by the collecting practices of academic institutions. One might argue that large-scale cooperative efforts are best organized within existing communities of interest: e.g., liberal arts colleges, research universities, or law schools. Indeed, one of the most successful approaches to scaling library service provision to a supra-regional level, the National Network of Libraries of Medicine (NN/LM) program, is explicitly limited to libraries serving the health services sector.[33]

It remains to be seen what benefits might be achieved if the scope of cooperative print management schemes were broadened to include a wider range of library types, and partnerships are established among institutions that have not collaborated in the past. The present study suggests that unless multi-type partnerships are established, it will be difficult—if not impossible—to ensure that the remarkable breadth and diversity of the North American print book collection is preserved for future citizens and scholars. It is possible that existing organizational structures will alter or expand their remit to enable them to represent the interests of other institutions in a common mega-region. But it is also possible that new organizations will emerge to fill the vacuum. This is essentially what has happened with the emergence of initiatives like WEST. Just as new cooperative structures have emerged to manage wildlife habitat, transportation systems, and high-speed computing networks on a regional scale, we can anticipate that management of library resources will increasingly require organizational structures that transcend existing political and institutional boundaries.

We do not claim that consolidation (physical or virtual) of print resources in North America should be organized at the mega-regional scale; rather, we have used the mega-regions framework to explore what a hypothetical regional consolidation of print resources might look like. Operationalizing a large-scale consolidation of regional print book collections will present many challenges; further study of the nature of these challenges and their potential solutions is needed. In particular, we hope this report will stimulate additional research on the factors that determine the appropriate scale of supra-institutional print management arrangements, as well as the cooperative and logistical infrastructure needed to sustain them.

Notes

1. Comparing discrete publications in HathiTrust against print book holdings in individual ARL libraries.
2. Subsequent analysis confirmed this projection. The slowed growth in overlap between 2010 and 2011 is partly explained by the evolving composition of the HathiTrust partnership and collection. The overlap will continue to fluctuate as a result of changing content contribution patterns (which affect the composition of the aggregated corpus), and changes in library acquisition trends (which alter the baseline against which overlap is calculated).

3. This strategy was examined at length for the journal literature in an analysis conducted by Ithaka S+R (Schonfeld 2011).
4. The present study does not address the relative preservation benefits of physical or virtual consolidation of print collections (Maniatis et al. 2005). More recently, Paul Conway and colleagues have examined a variety of utility-based metrics for assessing the quality of digital surrogates as a replacement for print materials (Conway 2011).
5. This is the model being explored by the Western Regional Storage Trust (WEST), which allows low- and moderate-risk titles in the archive to be shared under prevailing inter-lending rules.
6. For example, JSTOR has adopted a model of physical consolidation for its paper journal backfiles, utilizing the print repositories at California Digital Library and Harvard for this purpose. But a virtual model of consolidation is employed for JSTOR's rare or special collections, whereby the print originals are retained and managed by the organizations that own them (JSTOR 2012).
7. A 2009 study by Lavoie and Dempsey estimated that 14 percent of US-published print book titles in WorldCat were published prior to 1923 and therefore clearly in the public domain (2009). As of February 2012, OCLC Research analyses of the HathiTrust Digital Library collection indicate that about 1.15 million of the more than 5.16 million unique titles in the digitized collection—or about 22 percent—are in the public domain. These estimates exclude the large number of publications that might be classed as "orphan works," for which some copyright exceptions can be exercised. By some accounts, orphan works may account for as much as 50 percent of the digitized volumes in the HathiTrust collection (Wilkin 2011).
8. This visualization of the North American mega-regions, used here and in other graphics in this report, is based on figure 5 in Florida, et al. (2008, 470).
9. More specifically, we equate a "book" with a language-based monograph.
10. Readers familiar with the FRBR entity relationship model will recognize that a publication is equivalent to a FRBR manifestation, and a physical copy to a FRBR item.
11. Quarterly snapshots of WorldCat are maintained and programmatically enriched by OCLC Research to support a range of projects and prototypes. External researchers interested in making use of this data in their own studies are encouraged to contact OCLC Research, which can make provisions for access.
12. The authors thank Michelle Alexopoulos of the University of Toronto for arranging the provision of the mega-region ZIP/postal code data for our work.
13. In 2008, Mark Nelson predicted that many of the impediments to e-book adoption in academic libraries would be resolved within five years. In 2010, a survey of public library leaders found a high level of interest in e-book adoption but also pervasive concerns about restrictive licensing and platform interoperability (COSLA 2010). A recent report by the Pew Internet and American Life project finds that "the increasing availability of e-content is prompting some to read more than in the past and to prefer buying books to borrowing them" (Rainie et al., 2012). In the global consumer market, e-book adoption rates are already high and predicted to increase substantially (Bowker 2012).
14. Courant and Nielson's study examines print book storage costs under a variety of different circumstances, and concludes that space is the single greatest cost driver. The sheer physicality of print books limits options for cost-effective management (2010).
15. It should be noted that a print book publication's "rareness"—i.e., the fact that it is held by only a few institutions—does not necessarily imply that it is an exceptionally valuable contribution to the regional or system-wide print book resource. For example, its scarcity may owe to the obsolescence or low quality of its content.
16. The Pearson correlation coefficient is 0.46 for the twelve regions.
17. The US and Canadian extra-regional collections are the collective print book collections of all institutions located outside of the mega-regions in the US and Canada, respectively. We will say more about these collections later in the report.

18. The US and Canadian extra-regional collections are counted as "regions" in this result. For example, if a particular publication was available in BOS-WASH, CASCADIA, and several US locations outside the mega-regions, this would be counted as three "regions."
19. The OCLC Conspectus is a subject hierarchy, ranging from broad to specific subject descriptions. The analysis focuses on Conspectus divisions, which are broad disciplines of knowledge. Our analysis is confined to print book publications in WorldCat that have an assigned Conspectus division, which includes the majority of the publications in each regional collection (for more information, see OCLC 2012, sec. 1.2).
20. FAST (Faceted Application of Subject Terminology) is a streamlined, simplified version of the Library of Congress Subject Headings schema (for more information, see OCLC Research 2011).
21. Overlap computed using Whitehead (2012) comparison tool.
22. The PHOENIX anomaly may be due to the relatively high proportion of ARL holdings associated with its collection. Fifty-two percent of print book holdings in the region belong to ARLs, the highest percentage of any region. One might expect that ARLs would tend to have higher proportions of unique and diverse materials in their collections, as well as higher proportions of older materials.
23. CHAR-LANTA, a medium-sized collection, also is relatively young (25 years). An interesting outlier is NOR-CAL, which is the fourth largest collection but is tied with PHOENIX for the youngest collection with a median age of 23 years.
24. It should be noted that the largest regional collections are typically those where urban growth and associated educational and social infrastructure have built up over centuries of economic development. Thus, as one moves from East to West, the mega-region collections tend to decrease in size, scope and relative distinctiveness. To a significant extent, the size of regional collections is determined by historical factors that have shaped (and continue to alter) the identity of each mega-region. One might also argue that the diffuse nature of the "extra-regional" print book resource in North America is an artifact of social and cultural movements, including the institutionalization of land-grant universities and the library extension movement in the second half of the nineteenth century. Both of these contributed to the dispersion of library resources across geographic areas where urban development was limited. Additional study might show if the utilitarian orientation of these two movements has had a lasting effect on the complexion of the regional library collections. An overview of the Library Extension movement is provided in deGruyter (1980).
25. CHI-PITTS is the highest overlapping collection for BOS-WASH at 50 percent.
26. OCLC Research periodically compares digitized book titles in the HathiTrust collection to print book titles in WorldCat. An analysis of the HathiTrust collection as of late February 2012 identified about 4.9 million discrete book titles that could be mapped to print book titles (and holdings) in WorldCat.
27. Several state-based last copy policies, including those organized by CARLI in Illinois and by VALE in New Jersey, are identified in Malpas's *Shared Print Policy Review Report* (2009). The Center for Research Libraries (CRL) maintains a registry of print archiving efforts that includes several state- and province-level initiatives, including the Ontario Council of University Libraries (OCUL) Thunder Bay Agreement (CRL 2012).
28. For details on planning for a shared storage facility for academic libraries in Florida, see University of Florida 2010. For information about the Maine Shared Collections Strategy project, see MSCS 2011.
29. As of March 2012, there are 23 institutional content contributors to the HathiTrust Digital Library. One is located outside of North America. Of the remaining 22, about 25 percent are located in the BOS-WASH mega-region and an additional 40 percent are located in CHI-PITTS and CHAR-LANTA. If one looks at the relative contribution of content (digitized volumes) by mega-region, CHI-PITTS has a dominant presence, representing more than 50 percent of volumes in the aggregation. This is unsurprising, as the University of Michigan's contributions alone account for almost 45 percent of the total. The strong representation of NOR-CAL and SO-CAL content is explained by the very large

contributions sourced from the University of California system. Estimates presented here are based on published figures as of February 2012. HathiTrust Statistics on contributions to the HathiTrust collection are published monthly (HathiTrust Digital Library 2012).

30. As noted above, the figure of 4.9 million print book titles in HathiTrust is based on analysis of discrete book titles that could be identified in WorldCat as of March 2012. HathiTrust reported a total of 5.4 million book titles for the same period. The discrepancy in title counts reflect a difference in the way titles are disambiguated. Our analysis counts only those titles for which a discrete OCLC number corresponding to a print book title in WorldCat could be identified.
31. For a view of the growing market for services related to "managing down" print book collections in academic libraries, see Lugg and Fischer 2008, and Gilson and Strauch 2012. In January 2012, OCLC announced a strategic partnership with Sustainable Collection Services, LLC, an organization founded by Rick Lugg and Ruth Fischer to provide consulting and decision-support services to libraries reducing their print book holdings (OCLC 2012a).
32. The estimated cost of $30 per transaction is based on total-cost calculations established in 1993 (Roche 1993). Nearly two decades on, the costs have scarcely changed: a recent article cites costs ranging from $25 to $40 per volume (Esposito 2012).
33. The NN/LM includes more than six thousand health sciences libraries in the United States, encompassing university medical centers, teaching hospitals, and specialized research institutes. Network programs and services are coordinated by eight Regional Medical Libraries, which operate under contract to the National Library of Medicine. In 2011, the NN/LM initiated a regional print archiving program aimed at preserving at-risk medical journals (for more information about the MedPrint initiative, see NLM 2012).

References

Bowker. 2012. "Bowker Releases Results of Global eBook Research." About us: Press. 27 March. http://www.bowker.com/en-US/aboutus/press_room/2012/pr_03272012.shtml.

Conway, Paul. 2011. "Archival Quality and Long-Term Preservation: A Research Framework for Validating the Usefulness of Digital Surrogates." *Archival Science*. 11 (3-4): 293-309.

COPPUL (The Council of Prairie and Pacific University Libraries. 2011. "Fact Sheet." http://www.coppul.ca/keydocs/Fact_Sheet_11.pdf.

COSLA (Chief Officers of State Library Agencies). *2010. COSLA: eBook Feasibility Study for Public Libraries Final Report.* Portland, Oregon: Pinpoint Logic. http://www.cosla.org/documents/COSLA2270_Report_Final1.pdf.

Courant, Paul N., and Matthew "Buzzy" Nielsen. 2010. "On the Cost of Keeping a Book." In *The Idea of Order: Transforming Research Collections for 21st Century Scholarship*. (June): 81–105. Washington, DC: Council on Library and Information Resources. http://www.clir.org/pubs/abstract/pub147abst.html.

CRL (Center for Research Libraries). 2012. "Print Archives Registry." *Global Resources Network*. http://archivereg.crl.edu.

deGruyter, Lisa. 1980. "The History and Development of Rural Public Libraries." *Library Trends*. 28 (4): 513-23.

Esposito, Joe. 2012. "PDA and Inter-library Loan" *The Scholarly Kitchen* (blog). 13 March. http://scholarlykitchen.sspnet.org/2012/03/13/pda-and-inter-library-loan/.

Florida, Richard. 2008. *Who's Your City? How the Creative Economy is Making Where to Live the Most Important Decision of Your Life*. New York: Basic Books.

Florida, Richard, Tim Gulden, and Charlotta Mellander. 2008. "The Rise of the Mega-region." *Cambridge Journal of Regions, Economy and Society*. 1(3): 459-476.

Gammon, Julia, and Michael Zeoli. 2003. "Practical Cooperative Collecting for Consortia: Books-Not-Bought in Ohio." *Collection Management*. 28 (1/2): 77-105. http://www.crl.edu/sites/default/files/attachments/pages/5_0.pdf.

Gilson, Tom, and Katina Strauch. 2012. "ATG Interviews Rick Lugg and Ruth Fischer." *Against the Grain.* 24 (1): 34-36.

HathiTrust Digital Library. 2012. "Update on February 2012 Activities." About. News and Publications. All Newsletters and Reports. 9 March. http://www.hathitrust.org/updates_february2012.

Howard, Jennifer. 2011. "Building a Large-Scale Print-Journal Repository." Wired Campus (blog). *The Chronicle of Higher Education*. 8 February (5:33 pm). http://chronicle.com/blogs/wiredcampus/building-a-large-scale-print-journal-repository/29558.

ITHAKA. 2012. "Preserving Scholarship." JSTOR. Preservation: About. http://about.jstor.org/about-us/preserving-scholarship.

Lavoie, Brian, and Lorcan Dempsey. 2009. "Beyond 1923: Characteristics of Potentially In-Copyright Print Books in Library Collections." *D-Lib Magazine*. 15 (11/12). http://www.dlib.org/dlib/november09/lavoie/11lavoie.html.

Long, Matthew P., and Roger C. Schonfeld. 2010. *Ithaka S + R Library Survey 2010 Insights from U.S. Academic Library Directors*. New York: Ithaka S + R. http://www.ithaka.org/ithaka-s-r/research/ithaka-s-r-library-survey-2010/insights-from-us-academic-library-directors.pdf.

Lugg, Rick, and Ruth Fischer. 2008. "Future Tense—Weeding: The Time is Now." *Against the Grain*. 20 (4): 87-88. http://r2consulting.org/pdfs/2future_tense_lugg.pdf.

Malpas, Constance. 2011. *Cloud-sourcing Research Collections: Managing Print in the Mass-digitized Library Environment.* Dublin, Ohio: OCLC Research. http://www.oclc.org/research/publications/library/2011/2011-01.pdf.

Malpas, Constance. 2009. *Shared Print Policy Review Report*. Dublin, OH: OCLC Research. http://www.oclc.org/research/publications/library/2009/2009-03.pdf.

Maniatis, Petros, Mema Roussopoulos, TJ Giuli, David S. H. Rosenthal, and Mary Baker. 2005. "The LOCKSS Peer-to-Peer Digital Preservation System." *ACM Transactions on Computer Systems*. 23 (1): 2-50.

MSCS (Maine Shared Collections Strategy). 2011. "Collaborating to Preserve Our Print Collections." University of Maine. http://www.maineinfonet.net/mscs/.

Nelson, Mark. 2008. "E-Books in Higher Education: Nearing the End of the Era of Hype?" *EDUCAUSE Review* 43(2). http://www.educause.edu/ero/article/e-books-higher-education-nearing-end-era-hype.

Nitecki, Danuta, Carol Jones, and Jeffrey Barnett. 2009. "Borrow Direct: A Decade of a Sustained Quality Book-Lending Service." *Interlending and Document Supply*. 37 (4): 192-198.

NLM (US National Library of Medicine) "MedPrint—Medical Serials Print Preservation Program." *National Institutes of Health*. Last updated 01 March. http://www.nlm.nih.gov/psd/print_retention_about.html

OCLC. 2011. *Perceptions of Libraries, 2010: Context and Community*. Dublin, Ohio: OCLC. http://www.oclc.org/reports/2010perceptions/2010perceptions_all.pdf.

OCLC. 2012. "Introduction to the WorldCat Collection Analysis Service: The OCLC Conspectus." *WorldCat Collection Analysis User Guide*. Sec. 1.2. http://www.oclc.org/us/en/support/documentation/collectionanalysis/using/introduction/introduction.htm#conspectus_WCA.

———. 2012a. "OCLC Establishes Strategic Partnership with Sustainable Collection Services." *News Releases.* 20 January. http://www.oclc.org/news/releases/2012/20128.htm.

OCLC Research. 2011. "FAST (Faceted Application of Subject Terminology)" *Activities.* Last updated 18 November. http://www.oclc.org/research/activities/fast/.

Rainie, Lee, Kathryn Zickuhr, Kristen Purcell, Mary Madden and Joanna Brenner. 2012. *The Rise of E-reading*. Washington, D.C.: Pew Research Center's Internet & American Life Project. http://libraries.pewinternet.org/2012/04/04/the-rise-of-e-reading/.

Roche, Marilyn M. 1993. "ARL/RLG Interlibrary Loan Cost Study: A Joint Effort by the Association of Research Libraries and the Research Libraries Group." Washington, DC: The Association.

Schonfeld, Roger C. 2011. "What to Withdraw? Print Collection Management in the Wake of Digitization." *Serials Librarian*. 60 (1-4): 141-145.

University of Florida. 2010. "Shared Storage Facility for Library Materials." *George A Smathers Libraries*. Last updated 4 February. http://www.uflib.ufl.edu/communications/shared_storage_facility.html.

Whitehead (Whitehead Institute for Biomedical Research). 2012. "Compare Two Lists." *Bioinformatics & Research Computing Tools.* http://jura.wi.mit.edu/bioc/tools/compare.php.

Wikipedia. 2012. "List of Postal Codes in Canada." Last modified 6 May 2010. http://en.wikipedia.org/wiki/List_of_postal_codes_in_Canada.

——. 2012a. "List of ZIP Code Prefixes" Last modified 26 April 2012. http://en.wikipedia.org/wiki/List_of_ZIP_code_prefixes.

Wilkin, John P. 2011 "Bibliographic Indeterminacy and the Scale of Problems and Opportunities of 'Rights' in Digital Collection Building." *Ruminations*. February. http://www.clir.org/pubs/ruminations/01wilkin.

Subsidence and Uplift—the Library Landscape

Constance Malpas

There's been a lot of attention to geologic subsidence of late, what with all the sinkholes opening up in Florida, Louisiana and other places.[1] Here in California, we are more often concerned with the gradual change in ground level due to the draining of aquifers that support large-scale farming. From year to year, the difference in ground level may be nearly imperceptible but over the space of a few decades the landscape has been radically transformed.

The subsidence metaphor was on my mind recently, as I was looking over some data compiled by my colleague Thom Hickey,[2] documenting the usage of headings (subjects and names) in WorldCat. OCLC Research has done quite a lot of work exploring new approaches to managing subject and name authorities, notably in VIAF[3] and FAST.[4] I was interested to see how Thom's data might be used to measure change—uplift and subsidence—in the library landscape. By computing the frequency with which FAST and VIAF headings occur in institutional collections cataloged in WorldCat, one can identify which libraries hold the most materials related to particular topics, places and people. And this in turn provides a measure of the relative distinctiveness of library collections, judged not in terms of the "rarity" of holdings but rather by the concentration of related content.

It seemed to me that Thom's data might have something interesting to say about how the emergence of large-scale digitized book aggregations—HathiTrust,[5] Google Books,[6] etc—is altering the library environment. It stands to reason that as these large hubs begin to consolidate content sourced from libraries (and, in Google's case, publishers), they will displace traditional library "centers of excellence" in some subject areas. Those who remember the DLF Aquifer[7] project will recall that the initial prototype was designed to pool digitized resources in a given subject area (initially American History, later narrowed to Abraham Lincoln and the US Civil War). In the very large aggregations of HathiTrust and

This paper originally appeared in the OCLC Research Hangingtogether.org Blog on 18 April 2013. http://hangingtogether.org/?p=2680.

GoogleBooks, subject specialization has emerged more gradually. There has not been much public attention to measuring the scope of subject-based collections within those aggregations, nor to benchmarking them against existing institutional holdings.[8]

The FAST and VIAF centers data provide evidence of both subsidence and uplift in the current collections environment—that is, shifts in centers of excellence as measured by scope of subject based holdings. The "re-leveling" that has been wrought in just a few years of large-scale digitization is already significant. Digital aggregations have, by design or accident, emerged as important subject repositories that rival and even outrank some of the largest institutional libraries in WorldCat.

For instance, HathiTrust, an organization not yet five years old, already holds the greatest concentration of titles on the topic of marine biology,[9] surpassing the Library of Congress as well as two major research universities with world-class oceanography programs.

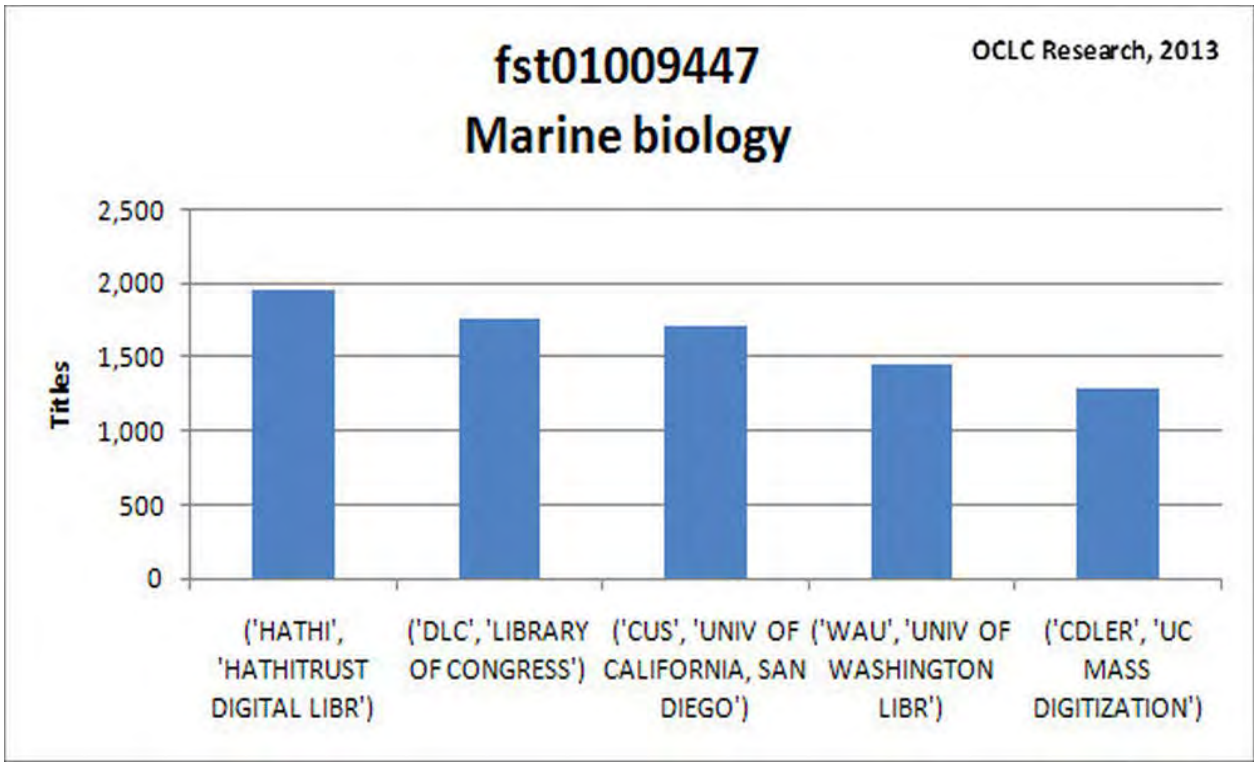


Figure 2. Number of marine biology tiles held in HathiTrust

In the case of Marine biology, the difference between the number of titles held by HathiTrust and the Library of Congress is not very large—fewer than 200 titles. But in other instances, the relative subsidence of traditional centers of excellence is more dramatic. For instance, Google Books substantially outranks several major research libraries in holdings related to Russian periodicals[10] (journals, newspapers and the like) (see figure 3).

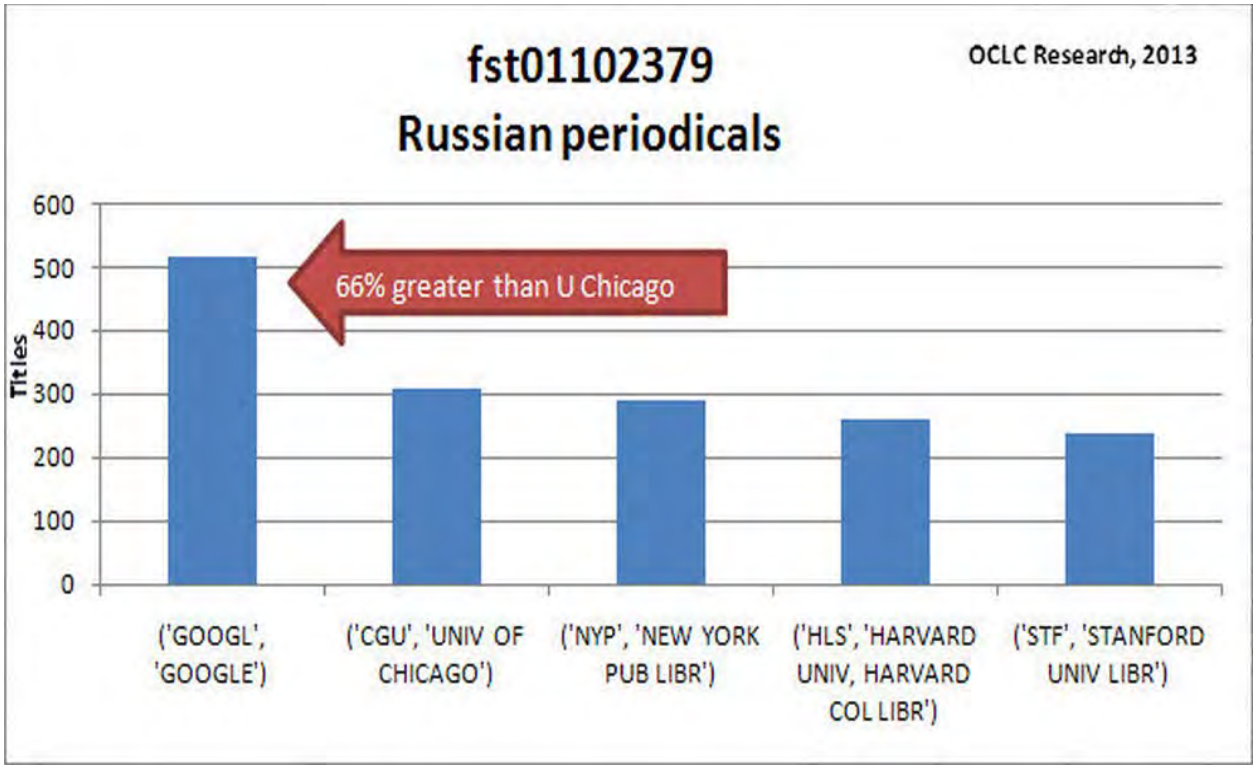


Figure 3. Number of Russian periodicals held in HathiTrust.

This represents an important change in the library system, with monumental old hubs being progressively overshadowed by new collections that are produced not by the slow accretion of library acquisitions but by large-scale digitization and (re)aggregation. It provides a compelling illustration of how Web-scale content aggregations are altering the library operating environment. In the case of HathiTrust especially, this disruption can (and I think should) be seen as a positive change: it enables libraries to rethink traditional, institution-scale collection management and stewardship—a topic we examined in our Cloud-sourcing Research Collections[11] report some years ago.

Using Thom's "centers" data, we can identify hundreds of topics and identities for which HathiTrust offers better coverage than any other library in WorldCat. Here a few topics in which the Digital Library distinguishes itself:

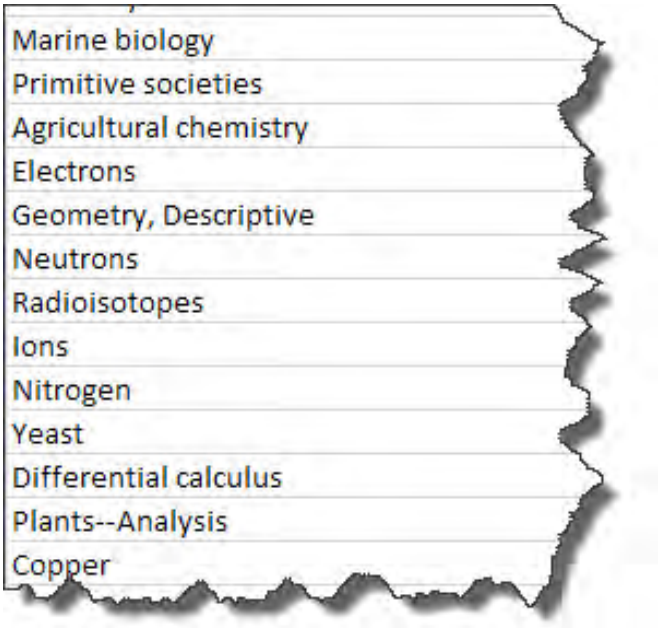


Figure 4. A sample of FAST subject headings for which HathiTrust holds more related titles (editions) than any other library in WorldCat.

And a few of the personal names for which its coverage is unrivaled:

Figure 5. A sample of VIAF personal name headings for which HathiTrust holds more related titles (editions) than any other library in WorldCat.

Interestingly, the other top-ranked collections (by size) for these same subjects and identities are not always the source of HathiTrust's richness. One might have anticipated that Hathi's leadership was simply a by-product of aggregating content from existing centers of excellence, but in fact Hathi has developed unexpected strengths by aggregating at a very large scale

from a diverse pool of contributors. For example, Harvard University and the University of Michigan each hold sizable collections of works by the poet Jean Ingelow; yet, the richness of Hathi's Ingelow collection is mostly due to contributions from campus libraries in the University of California system.

The FAST and VIAF "centers" data provide a fascinating new vantage point on the changing collections landscape. We'll be looking at ways to integrate it into ongoing research projects, including the mega-regions work,[12] where we hope it can help us detect regional collecting trends that might inform shared stewardship priorities.

Notes and References

1. Wine, Michael. 2013. "One Sinkhole Killed, and Many Others Opened, but Experts Counsel Not to Panic." *The New York Times. US edition*. 15 March. http://www.nytimes.com/2013/03/16/us/after-sinkhole-death-experts-say-unwarranted-frenzy-has-ensued.html?_r=2&.
2. See http://www.oclc.org/research/people/hickeyt.html.
3. See http://viaf.org/.
4. See http://www.oclc.org/research/activities/fast.html.
5. See http://www.hathitrust.org/home.
6. See http://books.google.com/.
7. Kott, Katherine, Jon Dunn, Martin Halbert, Leslie Johnston, Liz Milewicz and Sarah Shreeves. 2006. "Digital Library Federation (DLF) Aquifer Project." *D-Lib Magazine.* 12 (5). http://www.dlib.org/dlib/may06/kott/05kott.html.
8. Note: HathiTrust provides nice visualizations and a list of subject areas in the Digital Library, based on Library of Congress classification numbers. These provide a good overview of subject-based coverage but without reference to comparable coverage in other libraries. It is generally known that Google is selective with respect to identifying library partners, but I'm not aware of any public documentation related to a specific collection development strategy. Their aim, famously, is to provide comprehensive coverage of the world's books, not to develop excellence in any given subject area.
9. See http://experimental.worldcat.org/fast/1009447/.
10. See http://experimental.worldcat.org/fast/1102379/.
11. Malpas, Constance. 2011. Cloud-sourcing Research Collections: Managing Print in the Mass-digitized Library Environment. Dublin, Ohio: OCLC Research. https://www.oclc.org/content/dam/research/publications/library/2011/2011-01.pdf.
12. OCLC Research. 2013. "Print Management at 'Mega-scale': a Regional Perspective on Print Book Collections in North America." http://www.oclc.org/research/activities/megascale.html.

Contributors

Lynn Silipigni Connaway is a Senior Research Scientist at OCLC. She leads the OCLC Research User Behavior Studies & Synthesis activities theme. Her responsibilities include research projects that directly involve OCLC libraries and users, such as WorldCat data mining projects; JISC-funded investigations of digital information seekers, users in the virtual research environment, and—with the University of Oxford—digital "visitors" and "residents"; and IMLS-funded grant projects to study virtual reference services and the behavior patterns of college and university information seekers. Her complete bio is available online at http://www.oclc.org/research/people/connaway.html.

Lorcan Dempsey is Vice President, OCLC Research and Chief Strategist, OCLC. He oversees the research division and participates in planning at OCLC. He is a librarian who has worked for library and educational organizations in Ireland, England and the US. Lorcan has policy, research and service development experience, mostly in the area of networked information and digital libraries. He writes and speaks extensively, and can be followed on the web at Lorcan Dempsey's weblog and on twitter. Before moving to OCLC Lorcan worked for JISC in the UK, overseeing national information programs and services, and before that was Director of UKOLN a national UK research and policy unit at the University of Bath. Lorcan is Irish, and before moving to the UK he worked in public libraries in Dublin, Ireland. His complete bio is available online at http://www.oclc.org/research/people/dempsey.html.

Brian Lavoie is a Research Scientist at OCLC. Since joining OCLC Research in 1996, Brian has worked in a variety of research areas, including bibliographic control, analysis of library collections, models for library service provision, system-wide organization of library resources, analysis of the structure and content of the Web, and digital preservation. Brian is a co-founder of the award-winning PREMIS preservation metadata working group, and served on the PREMIS Editorial Committee. He co-chaired the international Blue Ribbon Task Force on Sustainable Digital Preservation and Access. His complete bio is available online at http://www.oclc.org/research/people/lavoie.html.

Constance Malpas is a Program Officer at OCLC. She works on issues related to measuring and shaping the outcomes of large-scale print conversion projects, and collaborative efforts to develop definitions and policies in shared print storage. Her complete bio is available online at http://www.oclc.org/research/people/malpas.html.

JD Shipengrover is a Senior Web and User Interface Designer at OCLC. Her primary focus is to bring user-centered interface design and usability principles to the Web applications created by OCLC Research. Her personal research interests include Web Usability, Mobile Design and Interactive Information Visualization. Her complete bio is available online at http://www.oclc.org/research/people/shipengrover.html.

Roger C. Schonfeld is Program Director for Libraries, Users, and Scholarly Practices. In this role, he leads Ithaka S+R's studies of academics' and students' attitudes, practices, and needs, as well research on the changing role of the academic library and scholarly society. He also consults with libraries and library consortia, digital humanities projects, distinctive collections and centers of excellence, and scholarly publishers. Roger has served on the NSF Blue Ribbon Task Force for Sustainable Digital Preservation and Access and NISO's Open Discovery Initiative. Earlier, he was a research associate at The Andrew W. Mellon Foundation, where he worked on projects related to college athletics and scholarly communication. Roger has a degree in English Literature from Yale University. http://www.sr.ithaka.org/people/roger-c-schonfeld-0.

Günter Waibel is Director of the Digitization Program Office at the Smithsonian Institution, where he oversees policy and strategy for digitizing and managing Smithsonian assets. He was previously a Program Officer for RLG and OCLC Research, where he focused on sharing, aggregating and disseminating cultural materials in a networked environment, and the intersection of libraries, archives and museums. He also served as Digital Media Developer at the UC Berkeley Art Museum & Pacific Film Archive and as Webmaster for the Oakland Museum of California. Günter served on the board of the Museum Computer Network (MCN) and the Association of American Museum's (AAM) Media & Technology Committee. He taught as adjunct faculty in the School of Information Studies at Syracuse University and currently teaches for the School of Library and Information Science at Catholic University of America.

For more information about our work related to understanding the system-wide library, please visit:
www.oclc.org/research/activities/usl.html

6565 Kilgour Place
Dublin, Ohio 43017-3395

T: 1-800-848-5878
T: +1-614-764-6000
F: +1-614-764-6096
www.oclc.org/research

ISBN: 1-55653-461-2
978-1-55653-461-4
MAN3185
1308/214959, OCLC