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

Constance Malpas

Program Officer
OCLC Research

OCLC™

A publication of OCLC Research

Cloud-sourcing Research Collections: Managing Print in the Mass-digitized Library Environment
Constance Malpas, for OCLC Research

Suggested citation:

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.

Contents

Figures

> 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.

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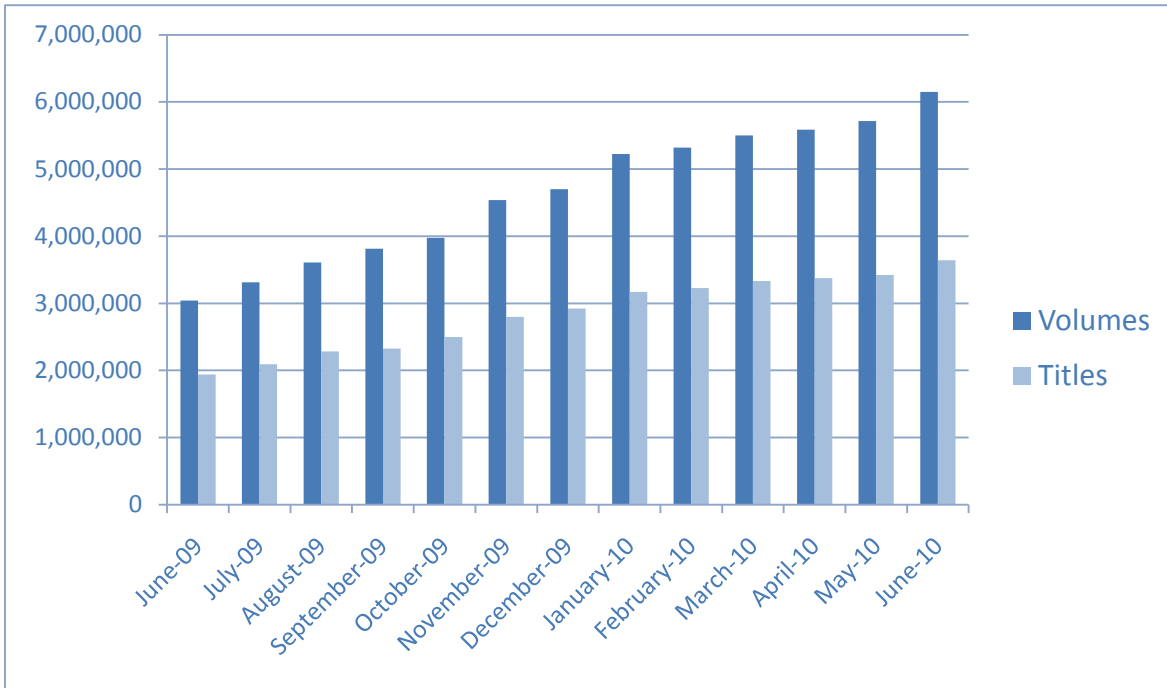
Figure 1. Growth of HathiTrust Digital Library collection (June 2009 - June 2010)

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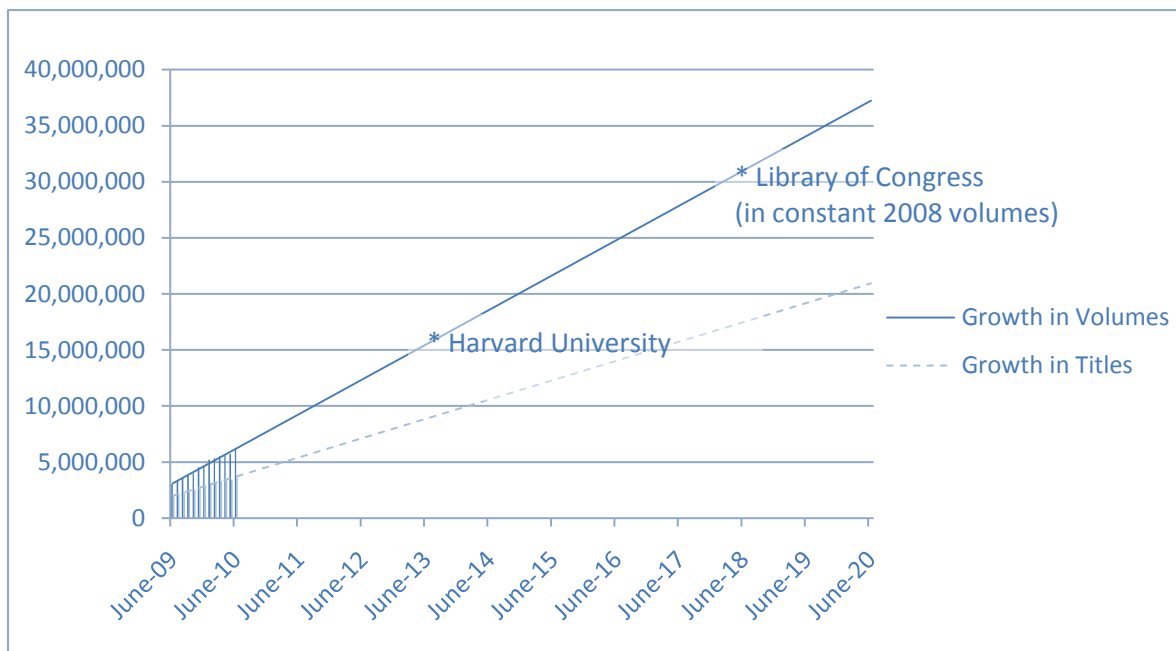
Figure 2. Projected growth of HathiTrust Digital Library (June 2010 - June 2020)

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.[i]

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 non-book and non-journal 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.

95%

- Books
- Serials
- Other

4%

1%

N = 3.64M titles
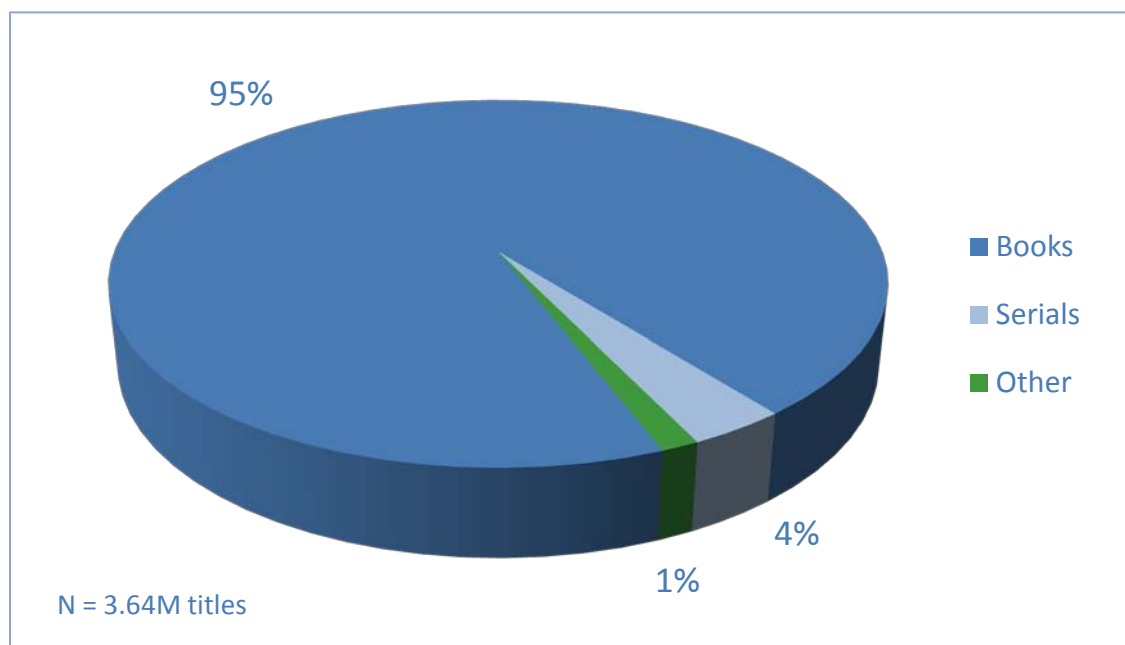
Figure 3. Primary document types of titles in HathiTrust Digital Library (June 2010)

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 non-book and non-serial 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 non-book 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 non-book 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.
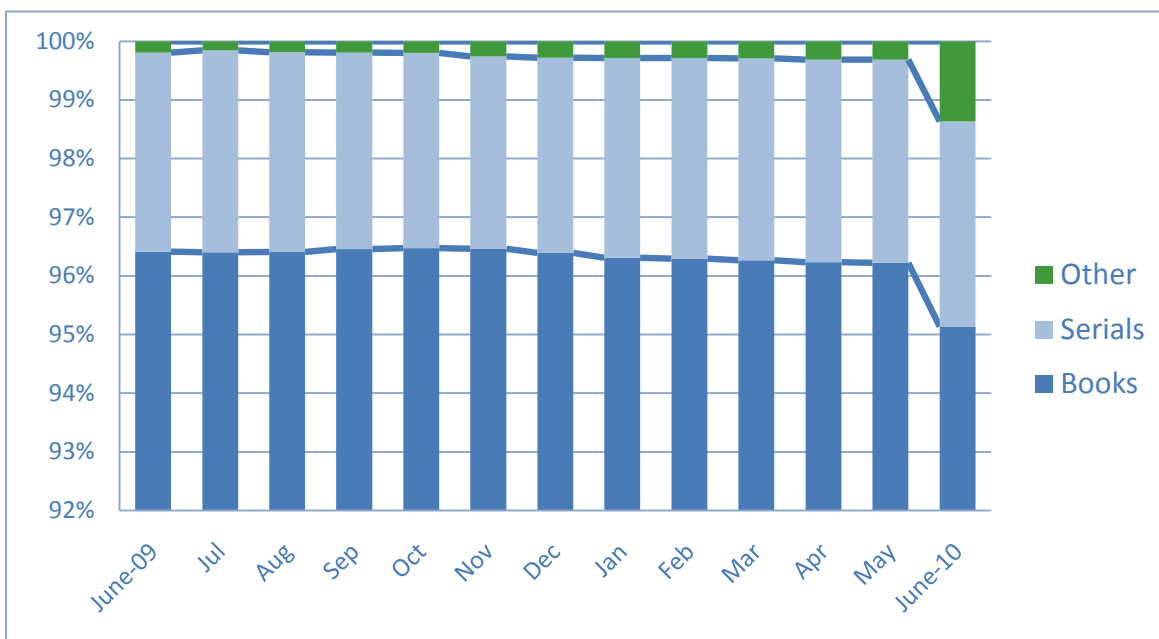
Figure 4. Distribution of HathiTrust Digital Library titles by document type (June 2009 - June 2010)

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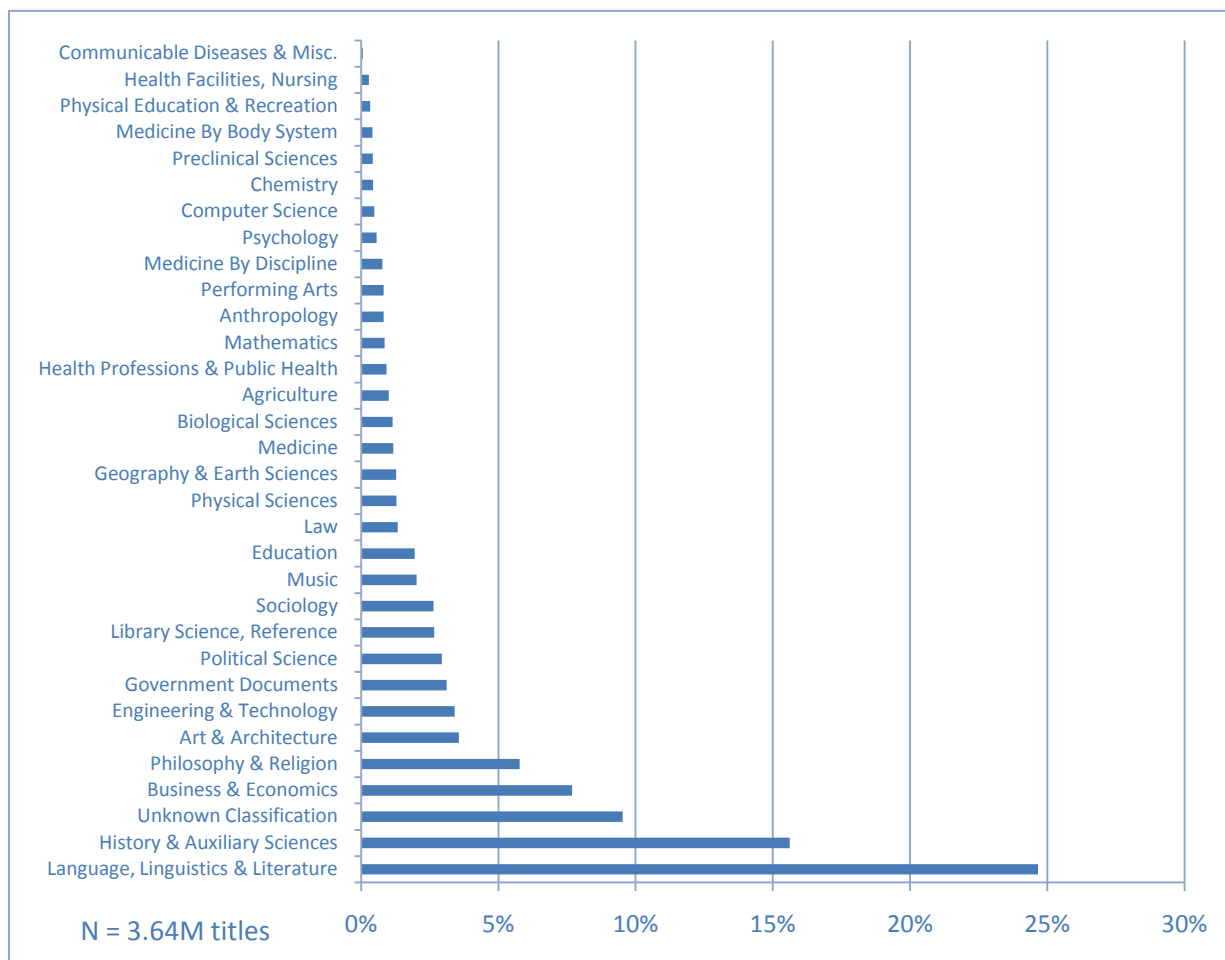
Figure 5. Subject distribution of titles in HathiTrust Digital Library (June 2010)

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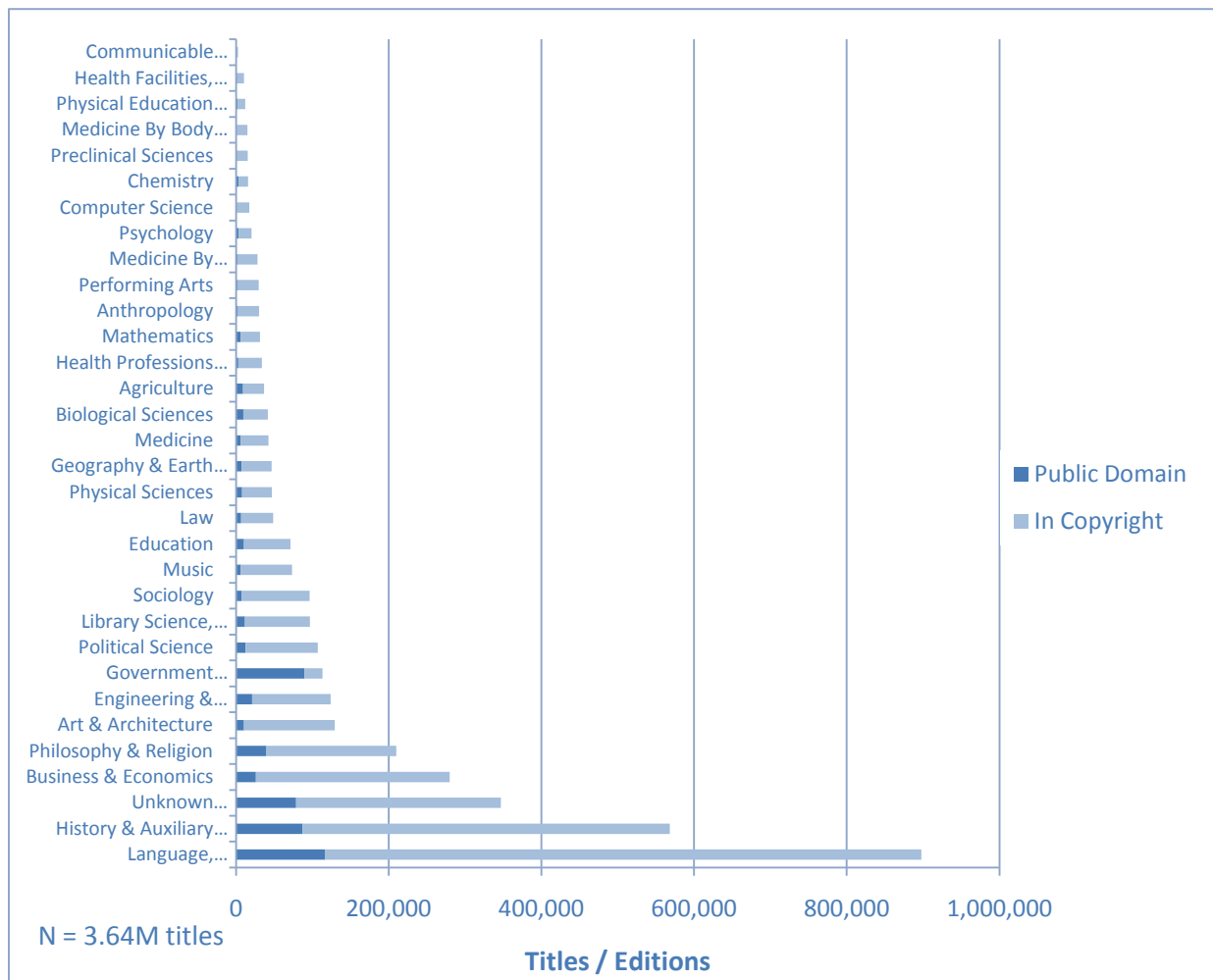
Figure 6. Distribution of titles in HathiTrust Digital Library by subject and copyright status (June 2010)

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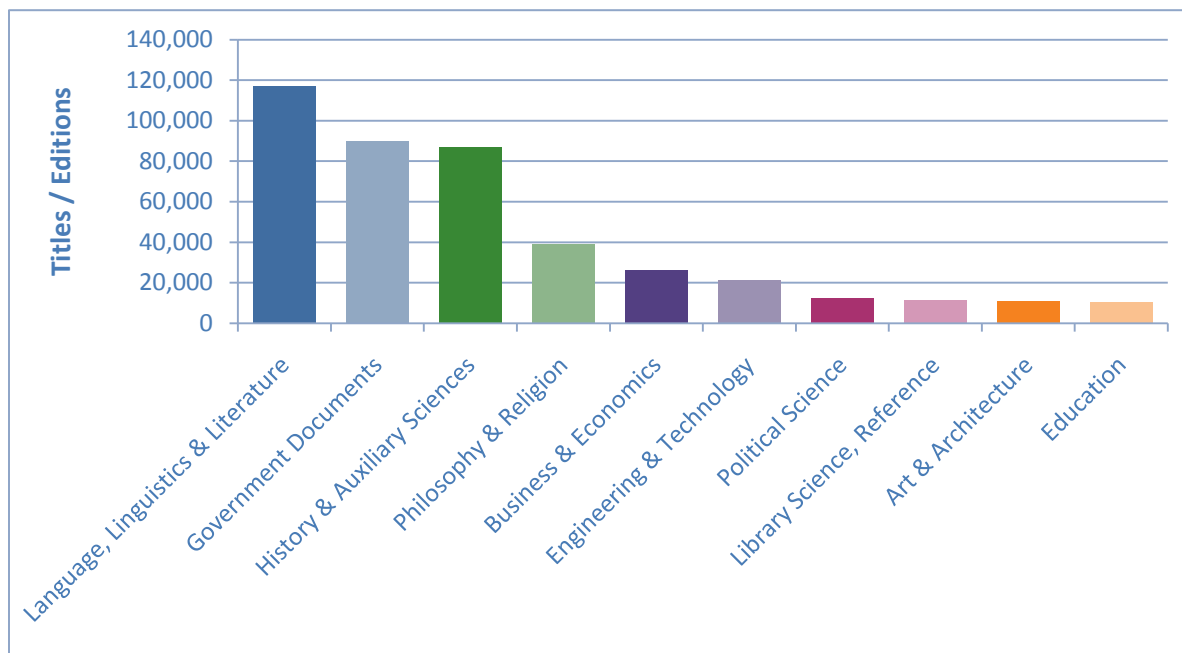
Figure 7. Top ten categories of public domain content in HathiTrust Digital Library (June 2010)

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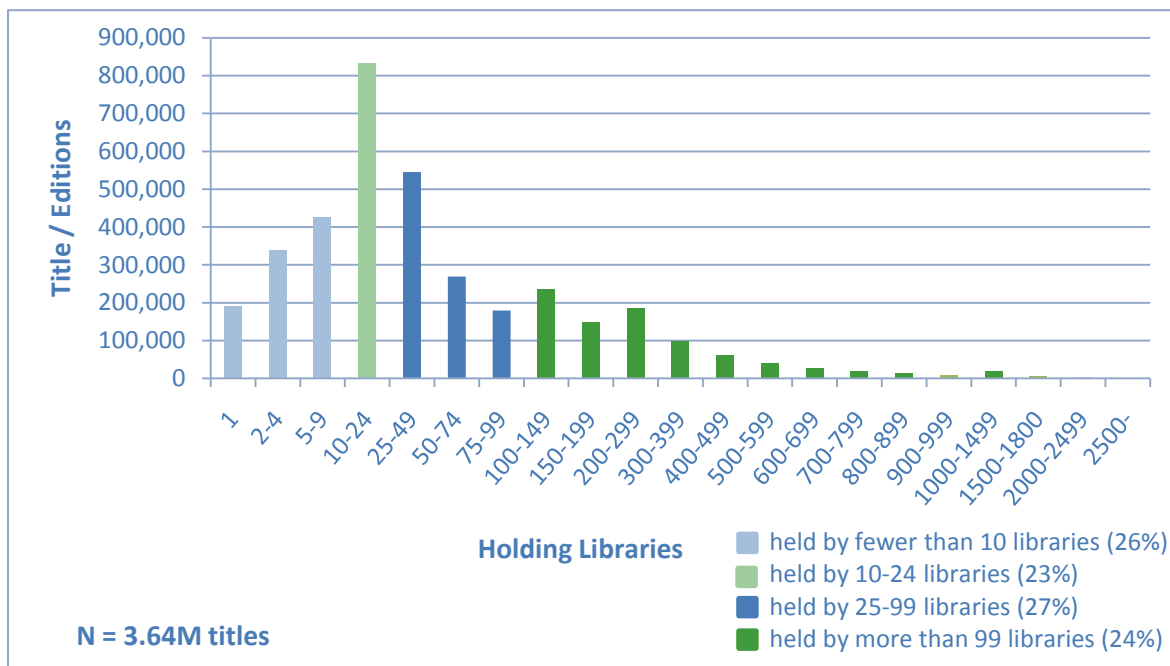
Figure 8. System-wide distribution of library holdings for titles in HathiTrust Digital Library (June 2010)

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 non-existent 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.

Princeton
University
Library
1,700,561
21%

Columbia
University
3,405,775
42%

The New York
Public Library
2,801,692
34%

Columbia
University Law
Library
263,330
3%
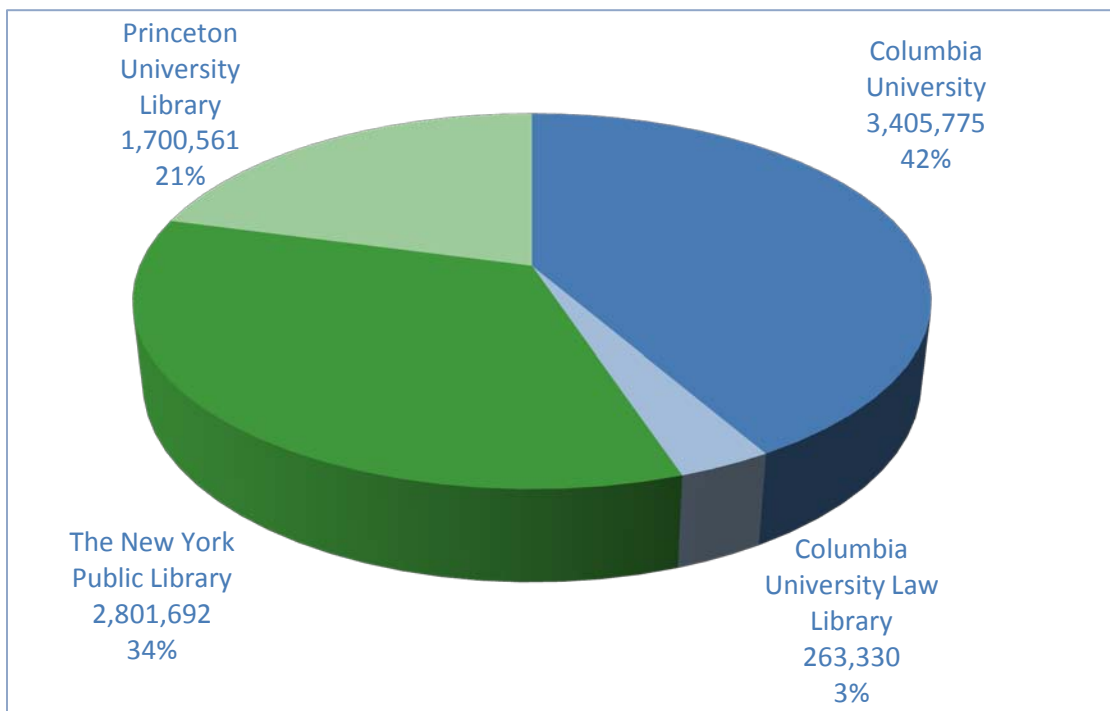
Figure 9. Distribution of ReCAP holdings by contributor (July 2010)

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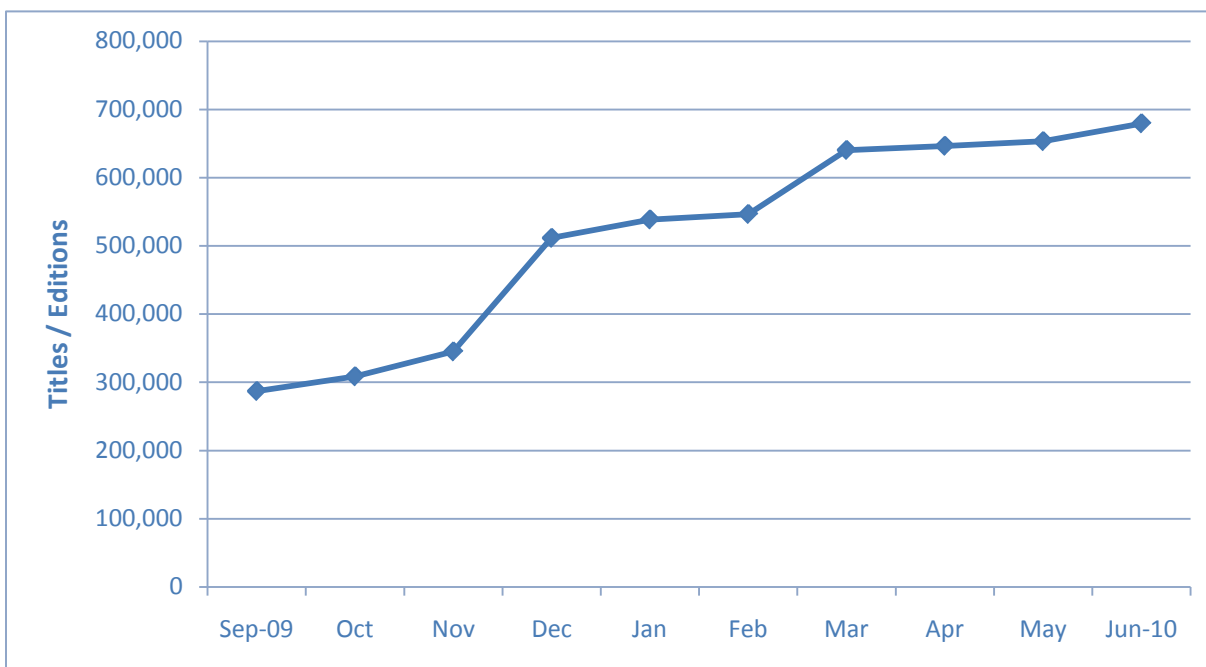
Figure 10. Growth in titles duplicated in ReCAP and HathiTrust Digital Library (September 2009 - June 2010)

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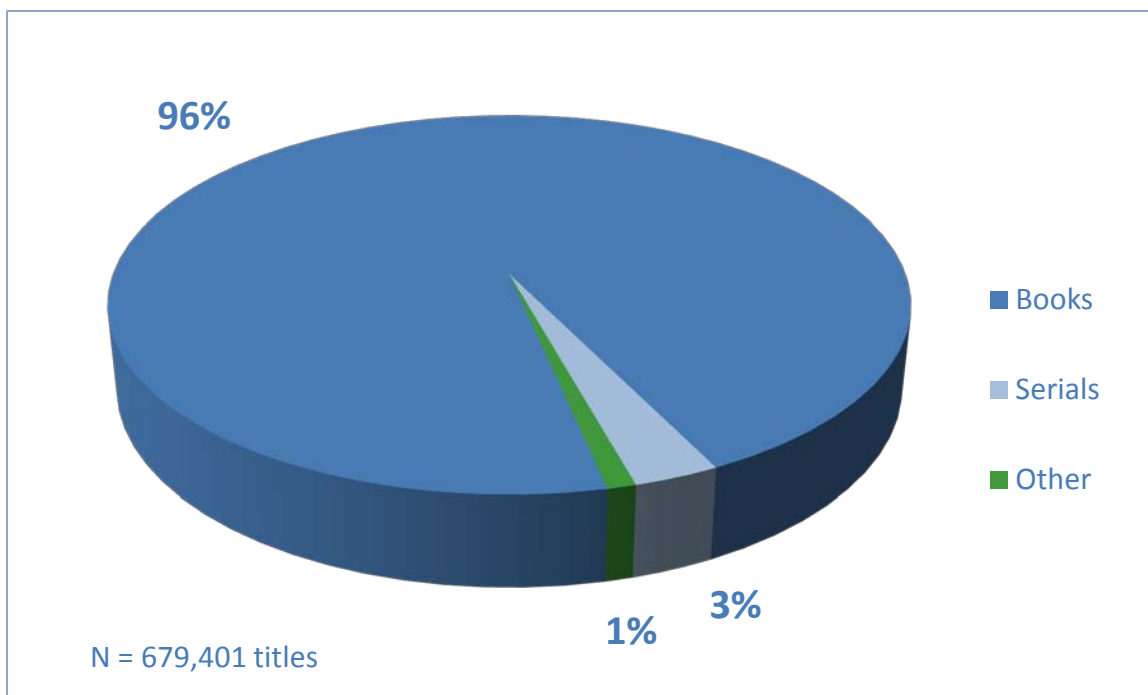
Figure 11. Primary document types of titles duplicated in ReCAP and HathiTrust Digital Library (June 2010)

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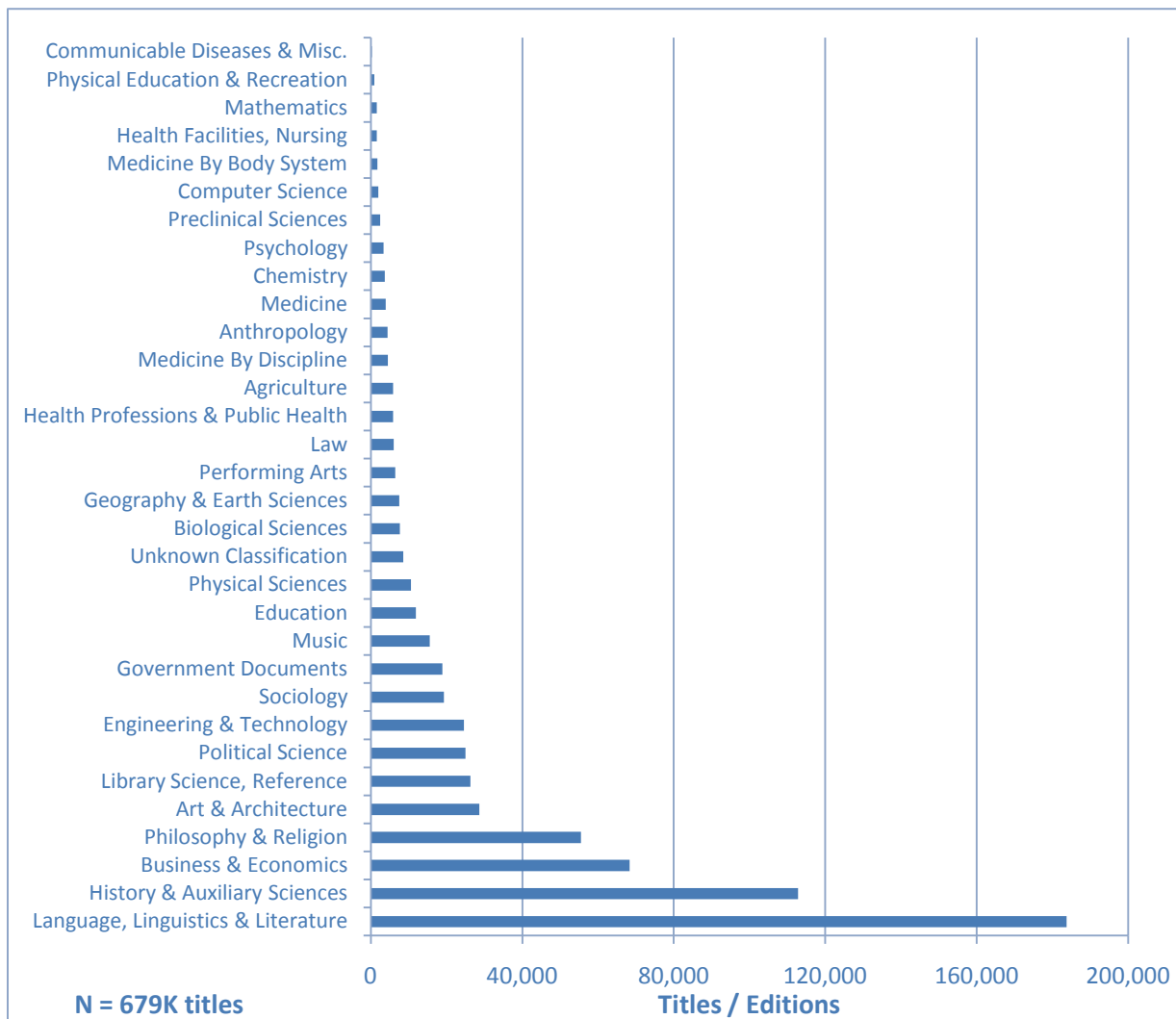
Figure 12. Subject distribution of Hathi titles held in ReCAP (June 2010)

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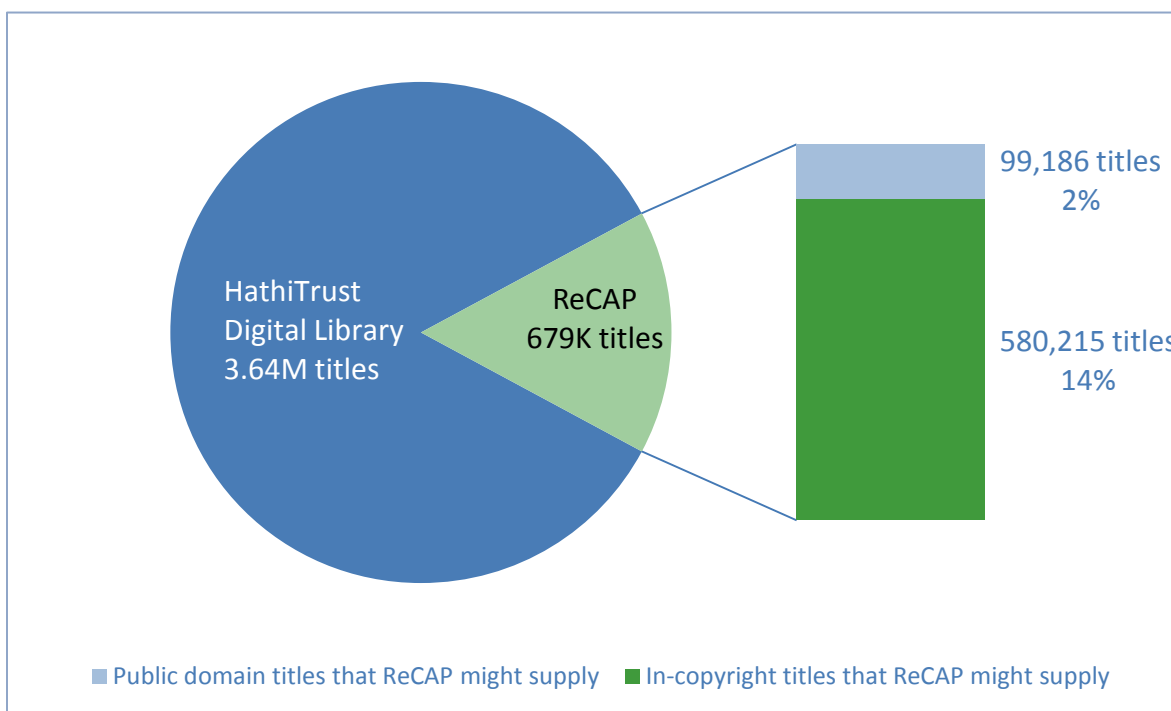
Figure 13. Comparative scope of shared digital and shared print repository collections (June 2010)

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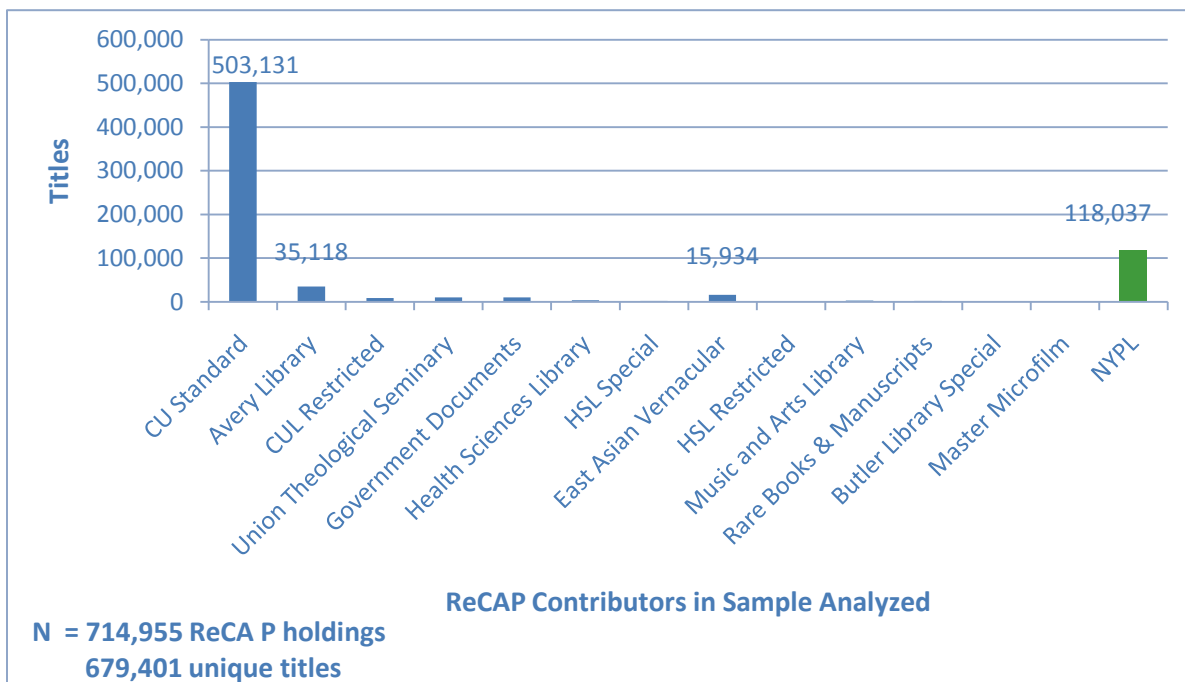
Figure 14. Titles duplicated in ReCAP and the HathiTrust Digital Library (June 2010)

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 seventy-plus 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 de-duplication 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 twenty-five thousand 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 de-duplicate and re-stock 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.
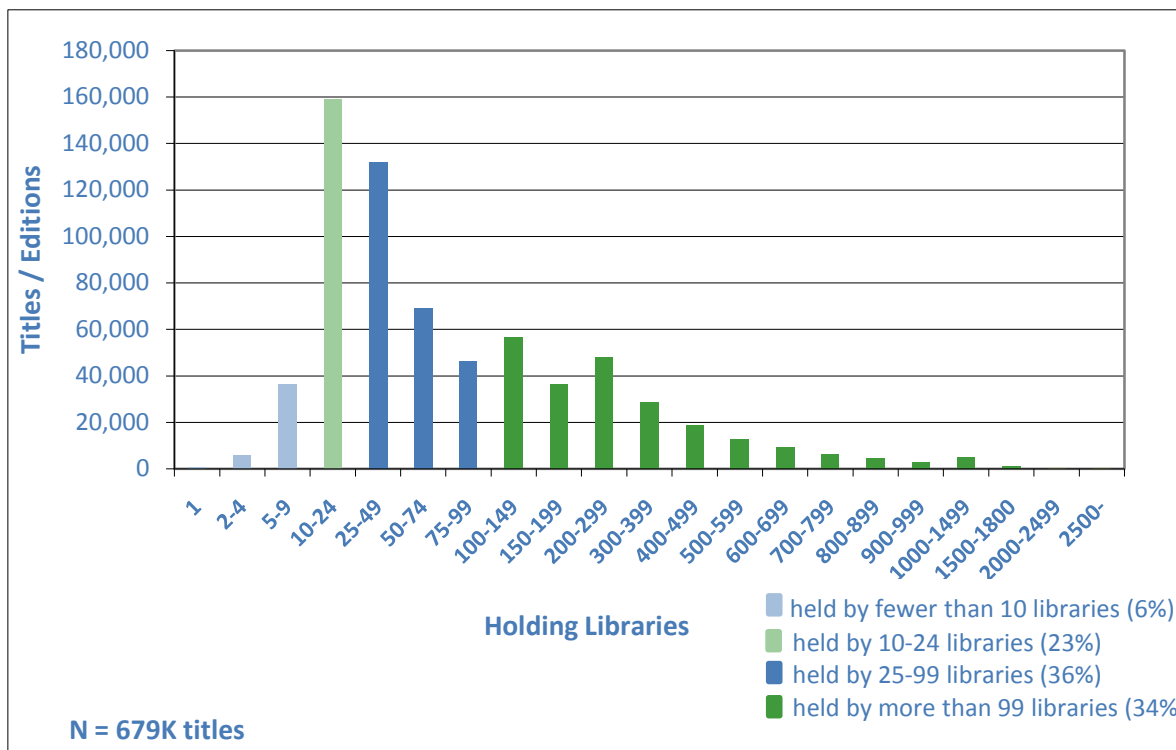
Figure 15. System-wide distribution of library holdings for Hathi titles in ReCAP (June 2010)

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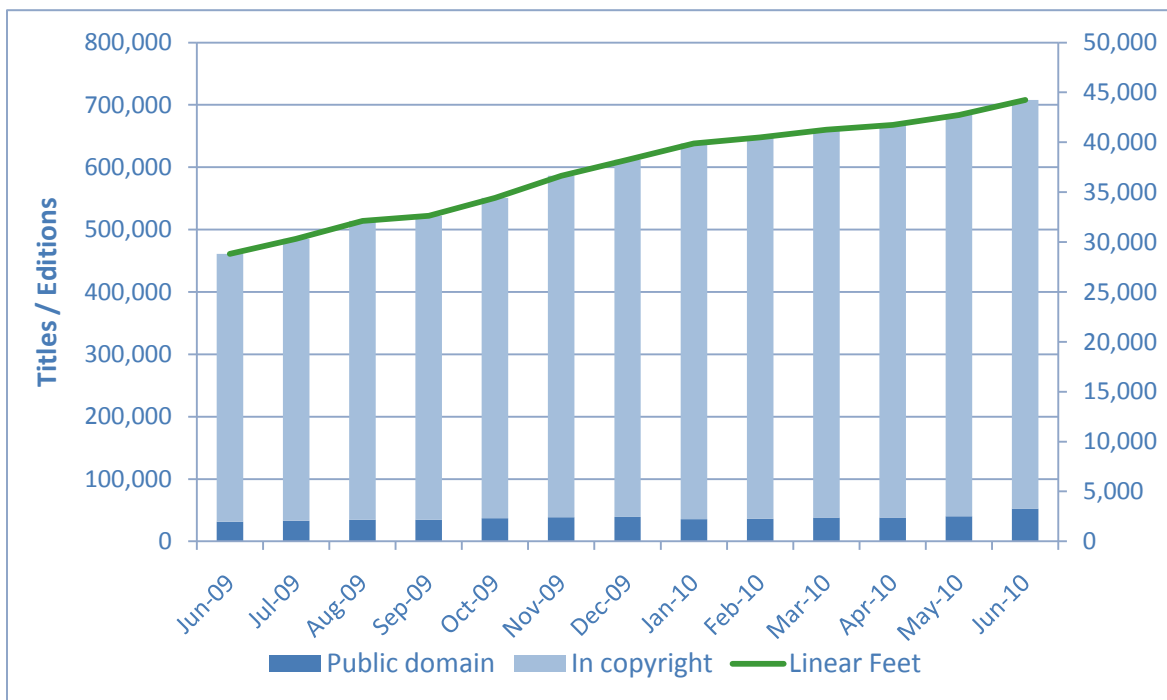
Figure 16. Growth in coverage of NYU Bobst holdings in HathiTrust Digital Library (June 2009 – June 2010)

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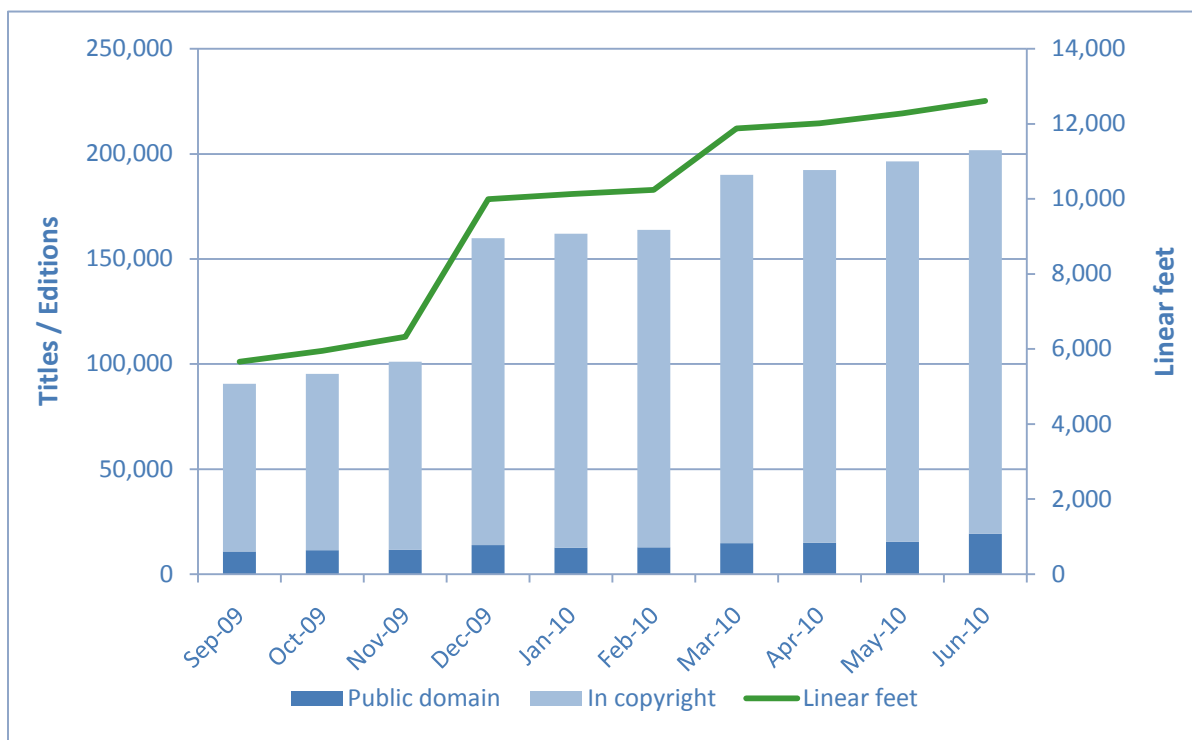
Figure 17. NYU Bobst titles duplicated in ReCAP and HathiTrust Digital Library (September 2009 – June 2010)

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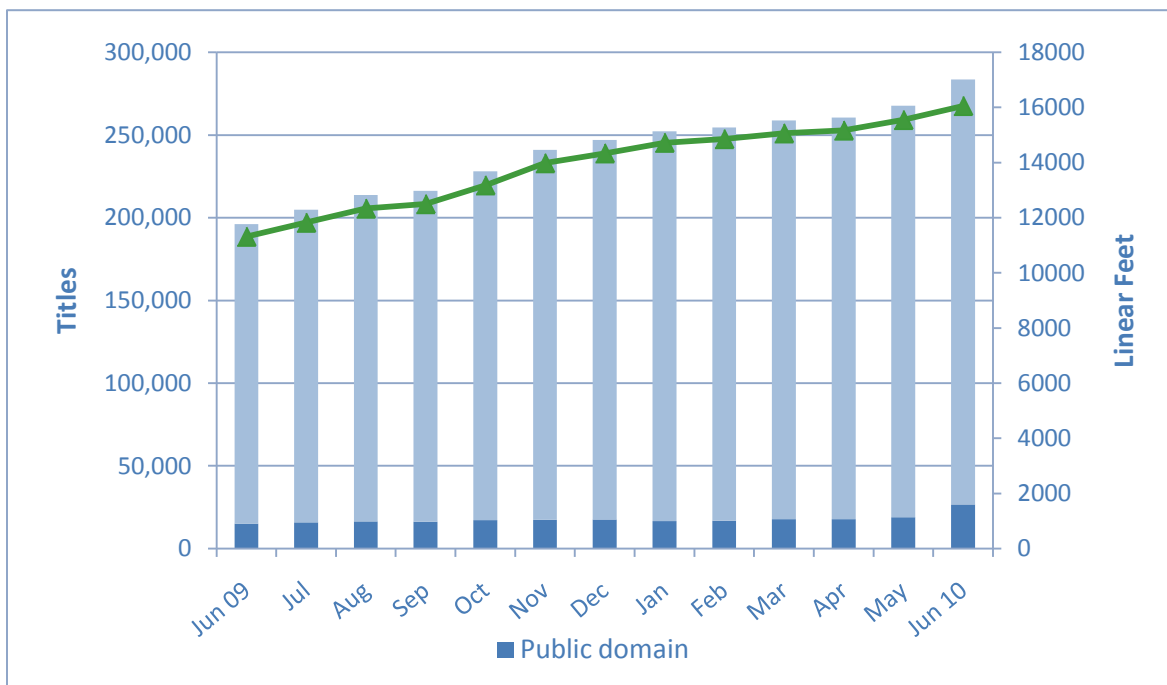
Figure 18. NYU Bobst titles duplicated in UC SRLF and HathiTrust Digital Library (June 2009 - June 2010)

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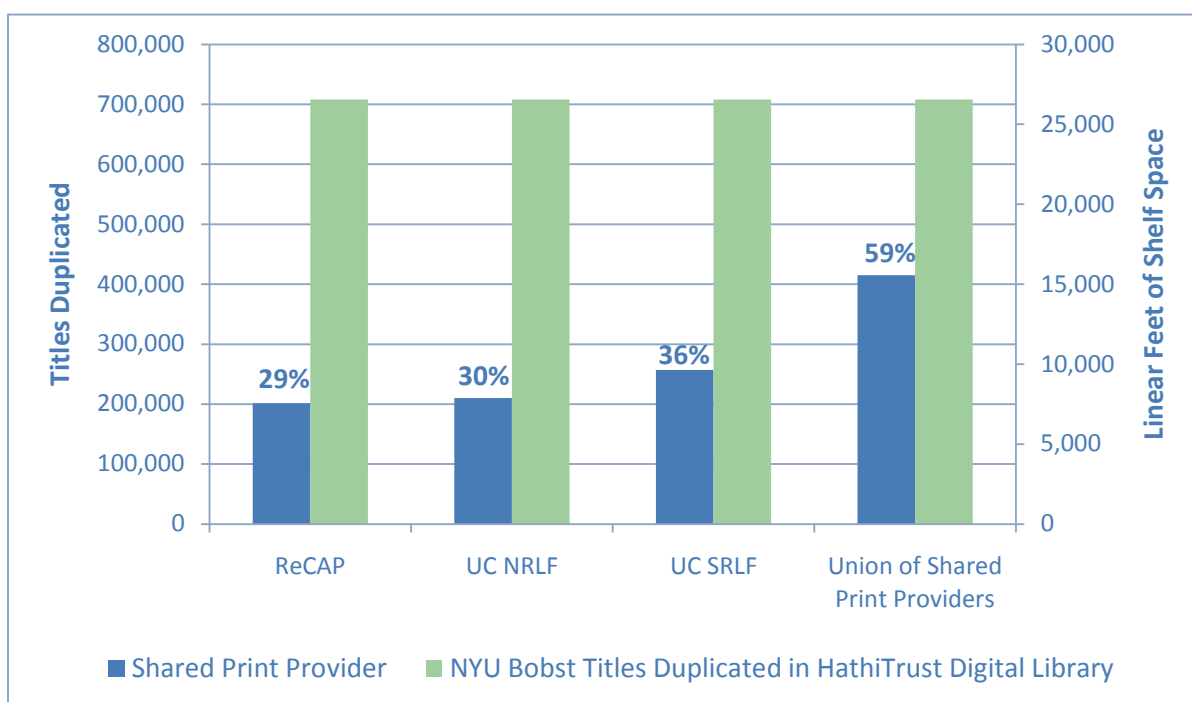
Figure 19. **Comparison of potential shared print provision options for NYU Bobst Library (June 2010)**

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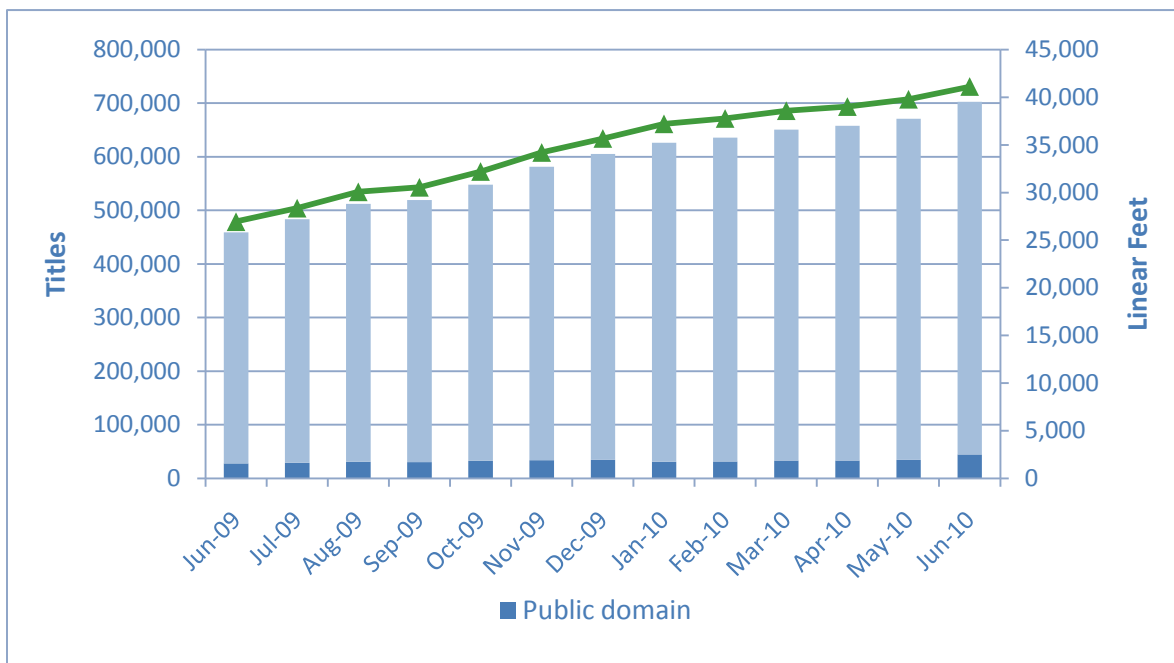
Figure 20. NYU Bobst titles duplicated in ReCAP partner libraries and HathiTrust Digital Library (June 2009 - June 2010)

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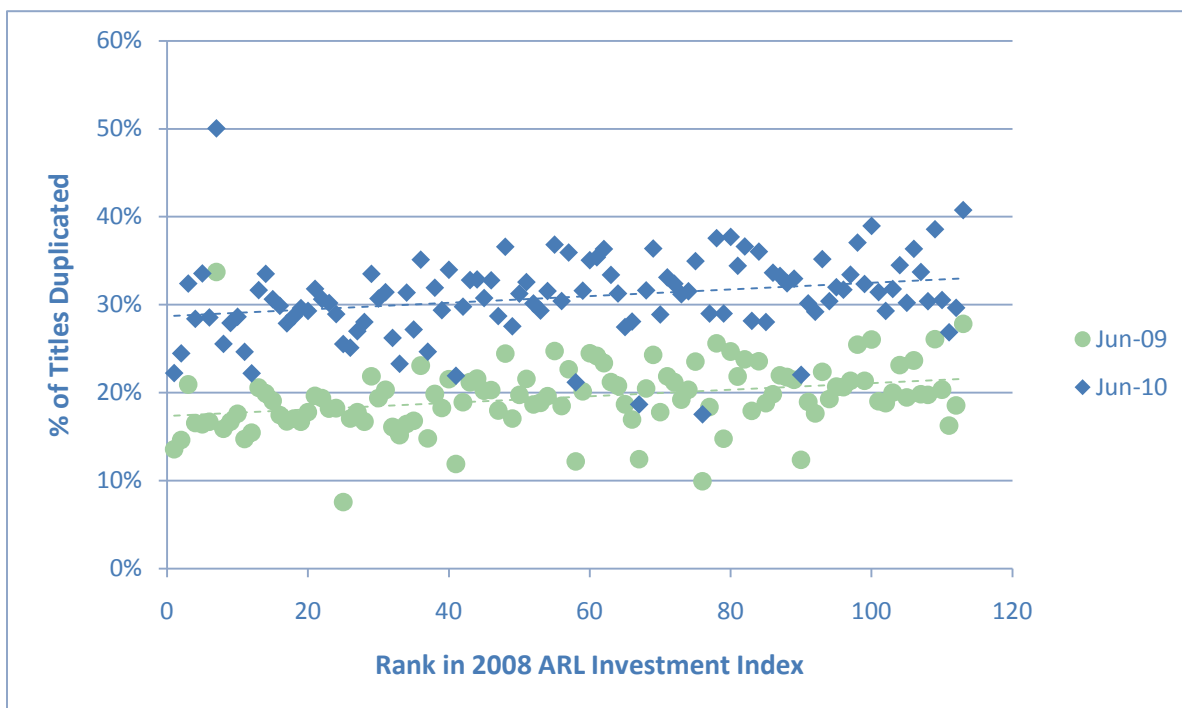
Figure 21. Percentage duplication of titles held in ARL libraries and HathiTrust Digital Library (June 2009 and June 2010)

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 increased reliance on digital surrogate. 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.

    o *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.

    o *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.

    o *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.

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)[1] 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.[2]

---

[1] 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.
[2] The philosophy of collective costs and collective holdings already underpins much of the CIC approach to HathiTrust.

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 re-calibrated 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

    o 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.

    o 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.

    o X (greater than one) is a value with which we multiply C to generate a surplus.

    o N is the total number of partner libraries.

> For a given in copyright volume, IC:
> $$IC = (C*X)/H$$

- where

    o C is the average annual cost to provide basic support for a volume (as above).

- o X (greater than one) is a value with which we multiply C to generate a surplus.

- o 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.

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. *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/.

---

[i] 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.

[ii] 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.