Stewarding the Collective Collection: An Analysis of Print Retention Data in the US and Canada

Ian Bogus
Research Collections and Preservation Consortium

Rachel Frick
OCLC

Devon Smith
OCLC

Susan Stearns
Eastern Academic Scholars' Trust (Retired)

Alison Wohlers
California Digital Library, University of California

OCLC®

November 2024

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

DOI: 10.25333/11f9-rw57
ISBN: 978-1-55653-296-2
OCLC Control Number: 1466216725

ORCID iDs
Rachel Frick: https://orcid.org/0000-0002-0184-4200
Devon Smith: https://orcid.org/0009-0002-1266-4579
Alison Wohlers: https://orcid.org/0009-0001-0804-4150

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

Suggested citation:
Bogus, Ian, Rachel Frick, Devon Smith, Susan Stearns, and Alison Wohlers. 2024. *Stewarding the Collective Collection: An Analysis of Print Retention Data in the US and Canada*. Dublin, OH: OCLC Research. https://doi.org/10.25333/11f9-rw57.

CONTENTS

FIGURES

INTRODUCTION

Shared print initiatives have been enormously successful. Libraries across the US and Canada have collectively agreed to retain millions of items and many shared print retention programs have reached significant milestones in their operational history. However, the print landscape is vast, and much remains to be done.

Accomplishing shared print's goals of protecting the print legacy will require coordination and a deep understanding of retention commitments. There has not been a comprehensive analysis of monograph shared print retention data since the genesis of shared print more than a decade ago. Without a comprehensive analysis across those commitments, and in comparison to the larger corpus of print monographs, the context needed to assess the impact of shared print and set priorities for moving forward is missing.

> **The research presented in this paper highlights the successes of shared print and identifies areas for strategic growth.**

To address this gap, OCLC Research and the Partnership for Shared Book Collections[1] (henceforth The Partnership) collaborated to design and execute a data analysis of monograph print retentions registered in the OCLC Shared Print Registry and situate those findings in the broader context of print monographs reflected in WorldCat. The research presented in this paper highlights the successes of shared print and identifies areas for strategic growth.

The points of examination were determined in consultation with the project partners. They were not intended to be exhaustive or explore other areas of interest beyond the current state of holdings data for shared print programs. As a result, the analysis provides baseline operational intelligence, mapping the current state of the shared print ecosystem and illuminating current risks and future opportunities.

Key findings from this analysis

1. **The vast majority of titles held in the United States and Canada do not have any copies with retention commitments, and for those that do, very few have redundant commitments.**

2. **Titles published between 1960 and 1990 have the greatest percentage of commitments, while more recent decades have significantly lower percentages of retention.**

3. **Subject area analysis based on Library of Congress (LC) Classification gives some insight, but more granular exploration is needed to ensure we are retaining inclusively.**

4. **Millions of items will reach the end of their retention period in the next five to 25 years.**

5. **Registered retention data is incomplete and can be hard to compare.**

These findings indicate potential action areas to strategically expand the shared print collection. More complete, comprehensive, and consistently available retention commitment data is crucial to concretely defining these strategic areas of expansion.

Background

Shared print is a combination of social and technical infrastructure that enables libraries to share responsibility for retaining print materials and providing access to them. It responds to a library collection paradigm shift from ownership to access. To ensure persistent access to titles, an adequate number of copies need to have commitments to overcome potential risks to the viability of retaining any individual copy. Shared print participants must determine their risk tolerance and the number of commitments needed to satisfy their tolerance.[2] They also need to collaborate to obtain commitments where there are too few or find other strategies when an adequate number of commitments cannot be made. Data about what is being retained and where it is being retained must be aggregated for analysis to make these decisions about the optimal number of commitments.

> **Shared print participants must determine their risk tolerance and the number of commitments needed to satisfy their tolerance.**

An essential part of the infrastructure that makes shared print valuable is making commitments public and visible—making a commitment *in camera* defeats the very purpose of the commitment. This disclosure of commitments allows, first, for the access that shared print is intended for and, second, for analysis that can identify gaps in the retained collection. In 2017, OCLC began its shared print registration service for monographs, and now, many programs require or encourage libraries to disclose their commitments through the OCLC registration service. Figure 1 shows these commitments are spread across the continent, with a strong concentration on the East Coast. There are other shared spaces where monograph shared print commitments are disclosed, such as program-specific registries, but the OCLC service is the most comprehensive for the US and Canada and includes:

- 14,101,512 titles with registered retention commitments
- 36,447,398 registered commitments to retain an item
- 352 libraries holding these registered titles
- 38 shared print programs across the US and Canada with registered commitments[3]

Map of registered print monograph retention commitments by region

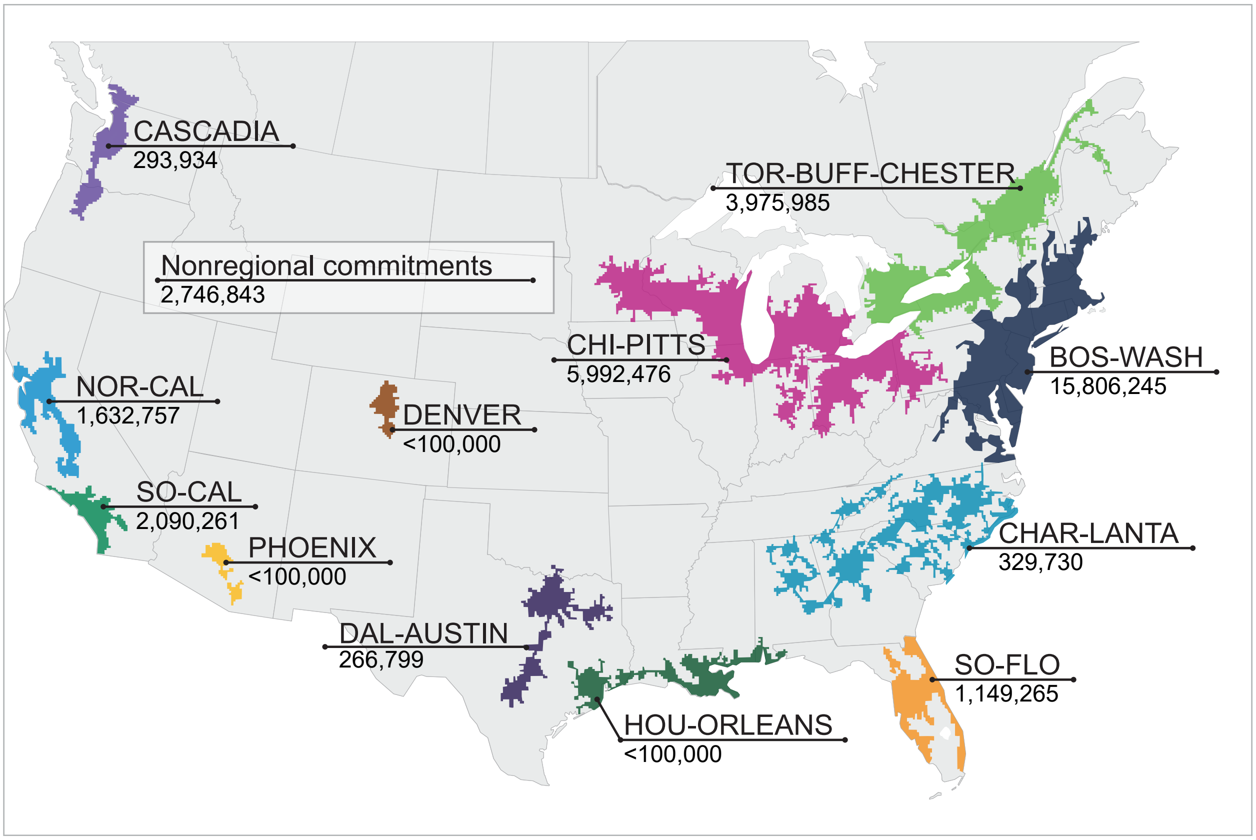


Figure 1. Map of registered print monograph retention commitments by region

It would be difficult to comprehensively understand the shared print landscape without registering all commitments across all programs. The absence of this information jeopardizes efforts to secure and make available our print legacy. The authors hope that by identifying where the shared print community needs more data and more commitments, this paper will catalyze and encourage more libraries to register their retention commitments.

Methodology

The shared print commitments and general holdings data under examination include local holding records (LHR) in MARC holdings format, WorldCat MARC bibliographic records, and an internal database of holding institutions. These data sets are combined to get the retention data from the LHRs, relevant bibliographic data from WorldCat, and location information from the holding institution database. The LHR table is the main linkage point. Each LHR contains an OCLC Control Number (OCN), which connects to WorldCat, and an identifier for the retaining institution, which connects to the institution table.

The LHR table contains many records beyond shared print retention commitments. To limit the set to just retention commitments, the LHR was filtered for records where the 583 $a contains the string "retain," case-insensitive. We also filtered for records where the Leader/Type Of Record is "x" (Single-part item holdings) or "v" (Multipart item holdings). The WorldCat institution table is global in scope, so it was filtered to remove institutions not located in the United States or Canada. WorldCat records were filtered so that only print monographs remained. When these filtered data sets are joined together, the final data set for examination represents registered retention commitments on single- or multipart print monographs held in the United States or Canada.

Data quality

Due to retention registration data standards, data quality procedures of the OCLC registration service, and the diligence of those submitting data to OCLC, the overall body of retention data is good, with more than 90% being in compliance and without significant issues.[4] Although the data quality is high, there is the challenge of unknown unknowns. Figure 2, which shows the overlap between institutions registering their commitments with OCLC and institutions participating in The Partnership (not all of which are retaining institutions), illustrates that a majority—65% (*N* = 320)—register with OCLC. However, 113 institutions in The Partnership either don't have retention commitments or haven't registered them with OCLC. We cannot comprehensively determine the total retention commitments because we have only the number of institutions not participating in the OCLC Shared Print registry and not their associated number of retention commitments. Additionally, not reflected in the data in the chart is the unknown number of institutions that have retention commitments but are both not in The Partnership and have not registered their commitments with OCLC. Despite this challenge, we are confident that the patterns observed based on 80% of the known universe would hold true against a total and complete data set.

Overlap between institutions registering their print monograph retention commitments with OCLC and institutions participating in The Partnership

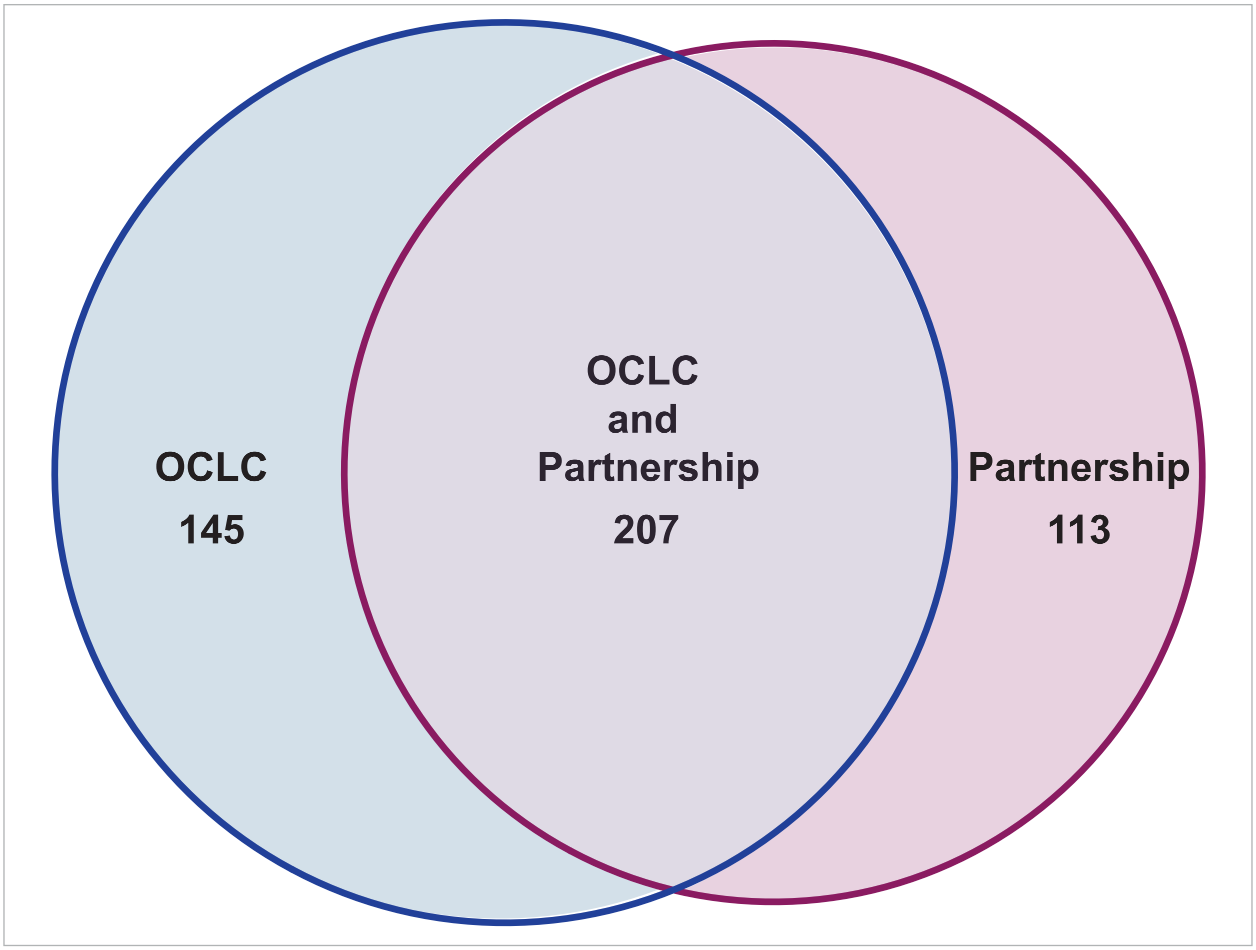


Figure 2. Overlap between institutions registering their print monograph retention commitments with OCLC and institutions participating in The Partnership

As with any analysis activity, there were challenges in preparing the data for analysis. Some of the challenges presented by the data may stem from the fact that bibliographic data is often cataloged using institutional local preferences stemming from their knowledge about their local user communities' needs. When data is aggregated for collective analysis, slightly different ways to express the information create impediments, if not real barriers, to conducting the analysis. As more attempts are made to understand the collective collection across increasing numbers of institutions, it will be more and more difficult to do so quickly and accurately. The community should consider this when planning any analysis that extends beyond their local collections and how they can "think global while acting local" in building their catalogs. Recording commitments that conform to standards and validating the data before making it public is the most important way to do this.

Analysis: What we found

This analysis evaluated retention commitment data based on regions, counts, programs, and other facets that help us understand the current state of the registered shared print universe. Establishing this baseline understanding will illuminate various trajectories that will maximize the benefit of shared print efforts.

US and Canadian regional coverage

Just over 100.5 million print monographic records in WorldCat are held by an institution in the United States or Canada. While shared print libraries have committed to retain a total of 36.4 million individual copies of monographs, they account for only 14.1 million WorldCat records. Of the 14% of monographs that have at least one copy with a retention commitment, 55% have only a single retention commitment (see figure 3). In other words, the vast majority of titles held in the United States and Canada do not have any copies with retention commitments. While it may seem that multiple commitments on a single title are wasteful, it is difficult to guarantee continued access to a single copy. Recent studies have shown that upwards of 3% of books that libraries think are on their shelves are missing from their collections.[5] Additionally, books may be damaged, incorrectly cataloged, or otherwise unusable.[6] Redundancy is key to preserving our written heritage and the ultimate success of shared print initiatives.

Percent of titles by number of commitments

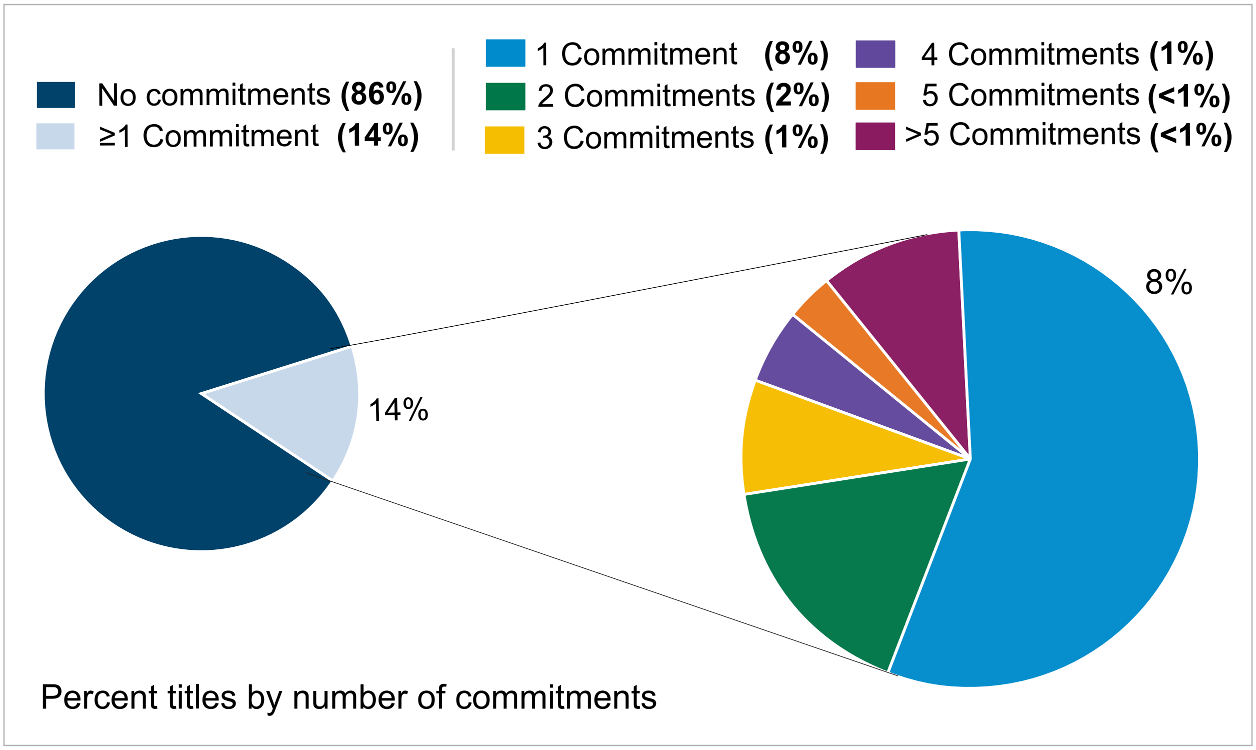


Figure 3. Percent of titles by number of commitments

Publication date coverage

There are about 95.2 million records of our original 100.5 million record set with valid dates between 1800 and 2020. As found in previous studies, the number of titles held in libraries increased dramatically starting in the 1950s.[7] The median publication decade was found to be the 1980s. The number of retention commitments follows a similar trend. We see a slightly different picture when we look at the percentage of records with retention commitments. Fewer than 10% of titles published in the first half of the nineteenth century have retention commitments (see figure 4). Titles published between 1960 and 1990 have the greatest percentage of commitments; about 17–18% of titles published during this period have commitments, significantly above the overall number. More recent decades have significantly lower percentages of retention. Only about 7% of the titles published in the 2010s have commitments, and only 1% of those published in the 2020s have commitments.

More work is needed to understand the possible relationship between items without shared print commitments and whether there are other types of intentions to retain.

The lower rate of coverage for older materials may be because some of those materials are in special collections. Including items held in restricted special collections can be seen as contrary to shared print's goals of improving accessibility. The disadvantage of not including them is that it may make it difficult to develop a complete picture of items that libraries intend to retain. More work is needed to understand the possible relationship between items without shared print commitments and whether there are other types of intentions to retain. There also may be ways to more easily identify items in special collections for intended retention.

While improvement in coverage is needed across the timeline, special attention is required for titles published in the nineteenth and twenty-first centuries. Many shared print programs have asked for commitments in waves from their member libraries. A more continuous method of setting commitments may alleviate the low percentage of commitments of recent publications.

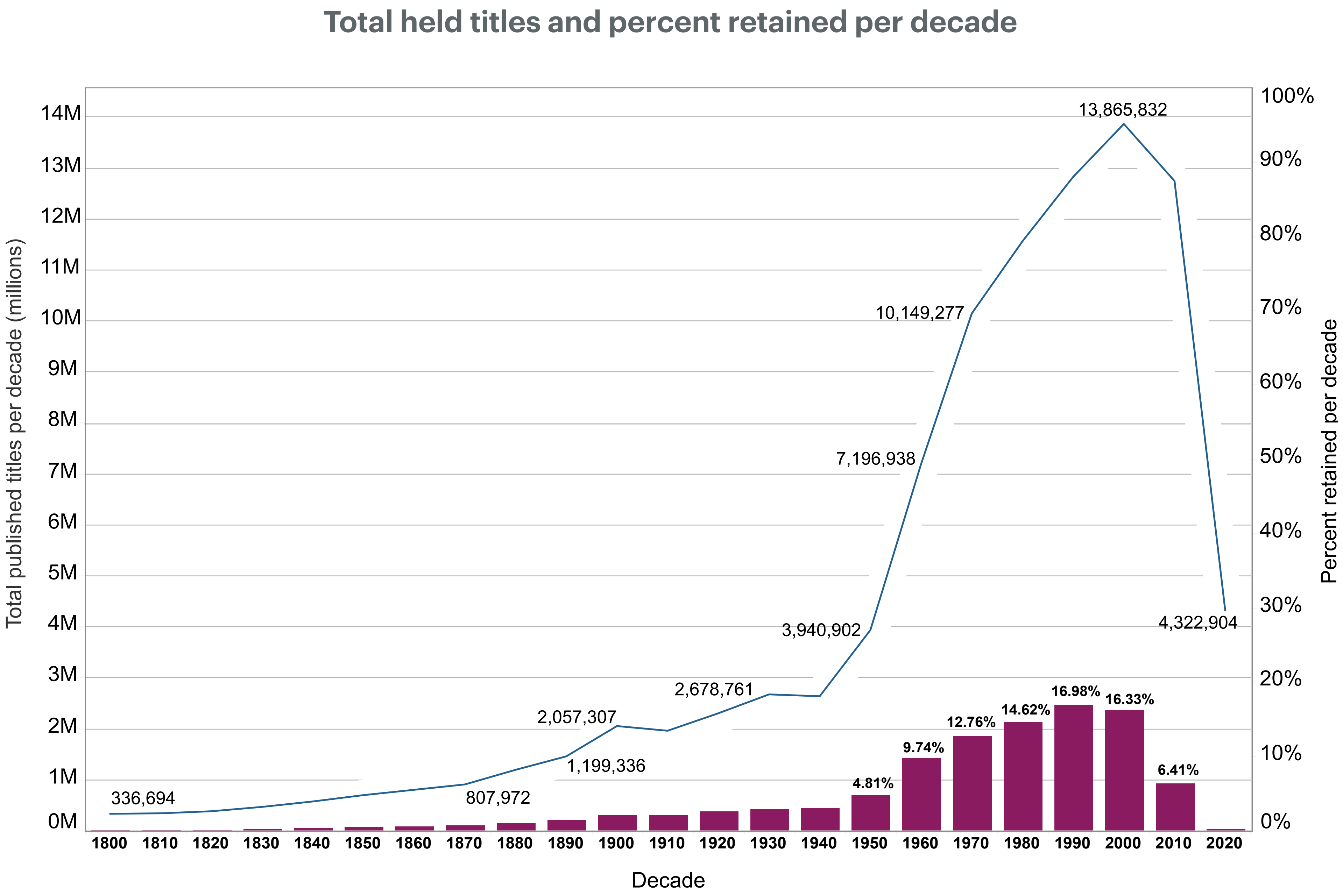


Figure 4. Total held titles and percent retained per decade

Subject coverage

To analyze the subject coverage of retention titles across programs and countries, Library of Congress Classification was used with a focus on the top eight classes represented in the data set[8]:

B - Philosophy, Psychology, and Religion

D - World History (excluding the Americas)

E - History of the Americas

H - Social Sciences

N - Fine Arts

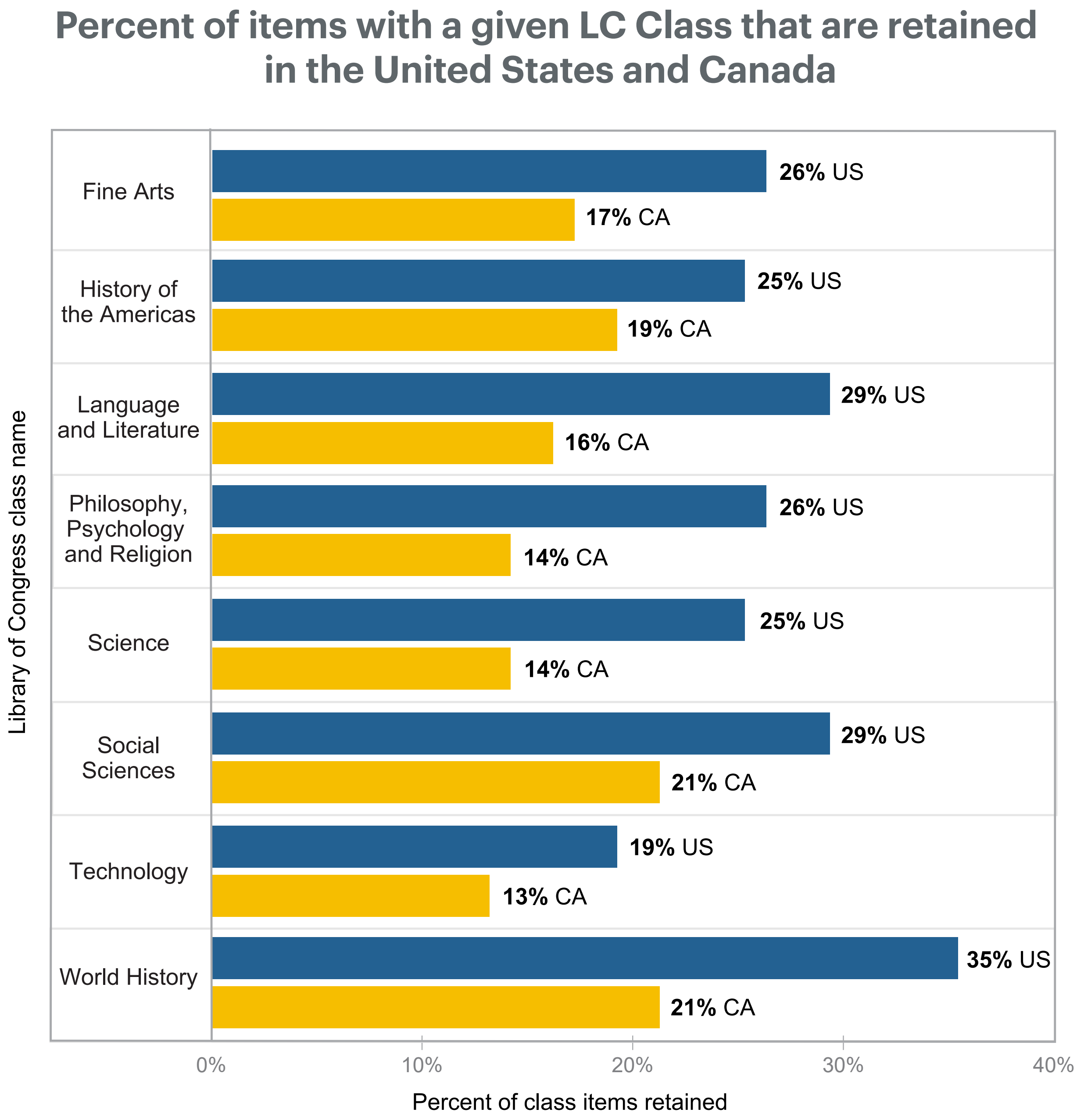
P - Language and Literature

Q - Science

T - Technology

While there were variations in subject coverage across the top programs in the US and between the American and Canadian programs, the overall trends were similar. The United States retains a somewhat larger percent of its holdings in each class, between 19–35%, while Canada retains between 12–21%. Technology is the lowest percent of retained items in both countries. World history is the largest percent of retained items in the United States, while in Canada, world history and social sciences tie for the largest percent of retained holdings.

While there were variations in subject coverage across the top programs in the US and between the American and Canadian programs, the overall trends were similar.



Percent of items with a given LC Class that are retained in the United States and Canada

Figure 5. Percent of items with a given LC Class that are retained in the United States and Canada

Drilling down to individual programs, the overall trends are similar, but some differences do appear. Comparing the four programs in the US with the largest number of retention commitments—EAST, HathiTrust, ReCAP, and SCELC—demonstrates that ReCAP is less technology- and science-rich as a percentage of total retentions, but it includes more world history than the other programs. SCELC retains the lowest percent of world history, but has the largest percent of philosophy, psychology and religion. All programs had a similar distribution of retention commitments across social sciences, fine arts, and language and literature, with language and literature being the largest percentage of all programs' retention commitments, as the chart below shows (see figure 6).

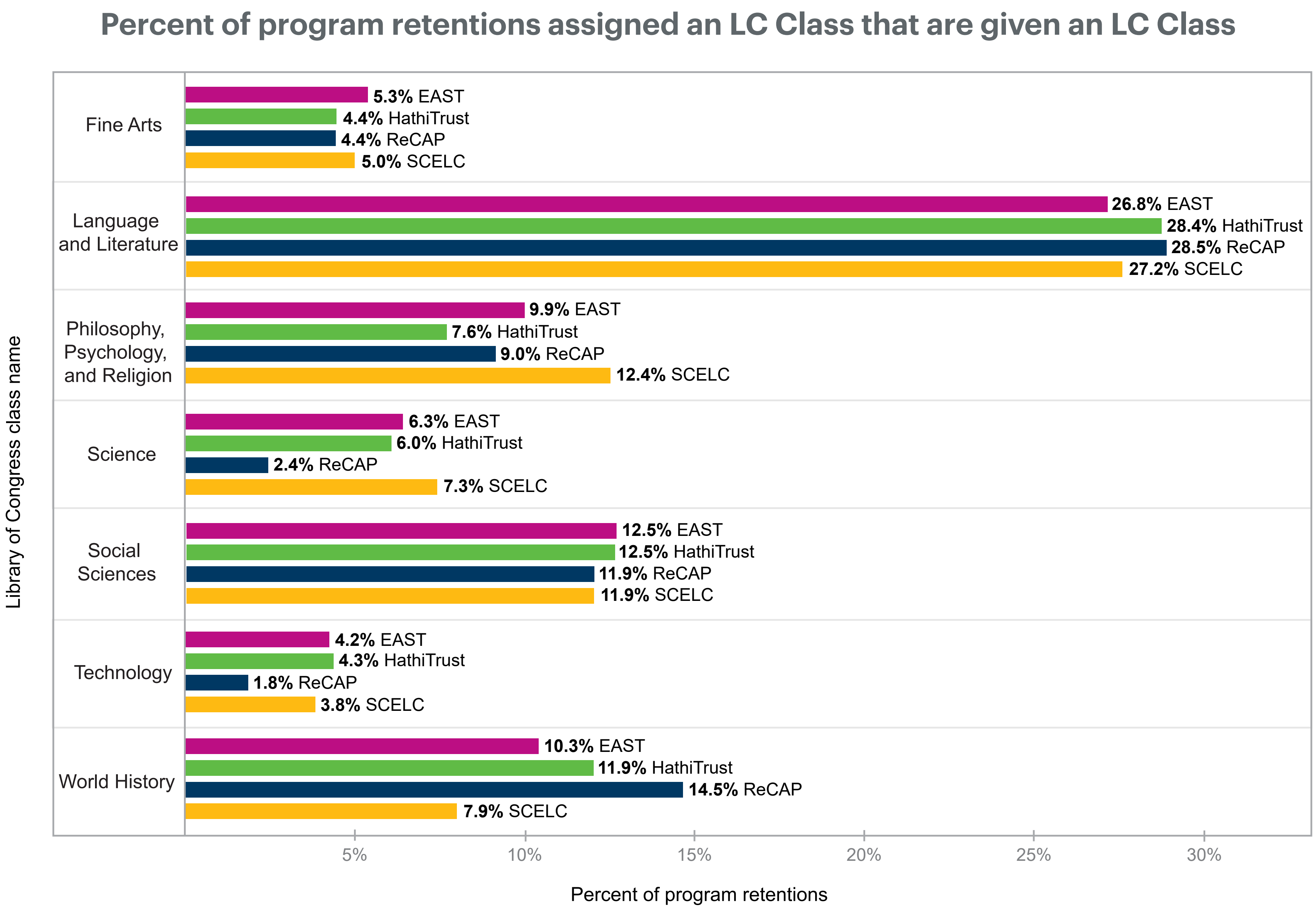


Figure 6. Percent of program retentions assigned an LC Class that are given an LC Class

This analysis does not illuminate subject areas that occur less frequently in retention commitments. More granular exploration, likely requiring analysis beyond simple comparisons by LC Class number, is needed to ensure that the most inclusive breadth and depth of the cultural and scholarly record is being actively stewarded in academic and research libraries in the US and Canada.

Risks, priorities, and calls to action

To best assess the paths that shared print programs may choose to prioritize their future efforts, it is important to identify current weaknesses, current operational practices and collections data, and other environmental risks associated with shared print programs more generally. For this data analysis, the following points were determined to be of immediate importance:

- Expiration dates
- Data quality and consistency

Community priorities: Expiration dates

While variations exist across the shared print programs analyzed here, the majority include some reference to when retention commitments "expire"—typically indicating when the shared print program participant is no longer responsible for either retaining the item, making it available, or both. This creates a pressing risk: Based on the current data available, close to 100% of the current retention commitments will expire within the next 25 years unless they are renewed in the interim (see figure 7).

This creates a pressing risk: Based on the current data available, close to 100% of the current retention commitments will expire within the next 25 years unless they are renewed in the interim.

Timeline of commitment expiration by program

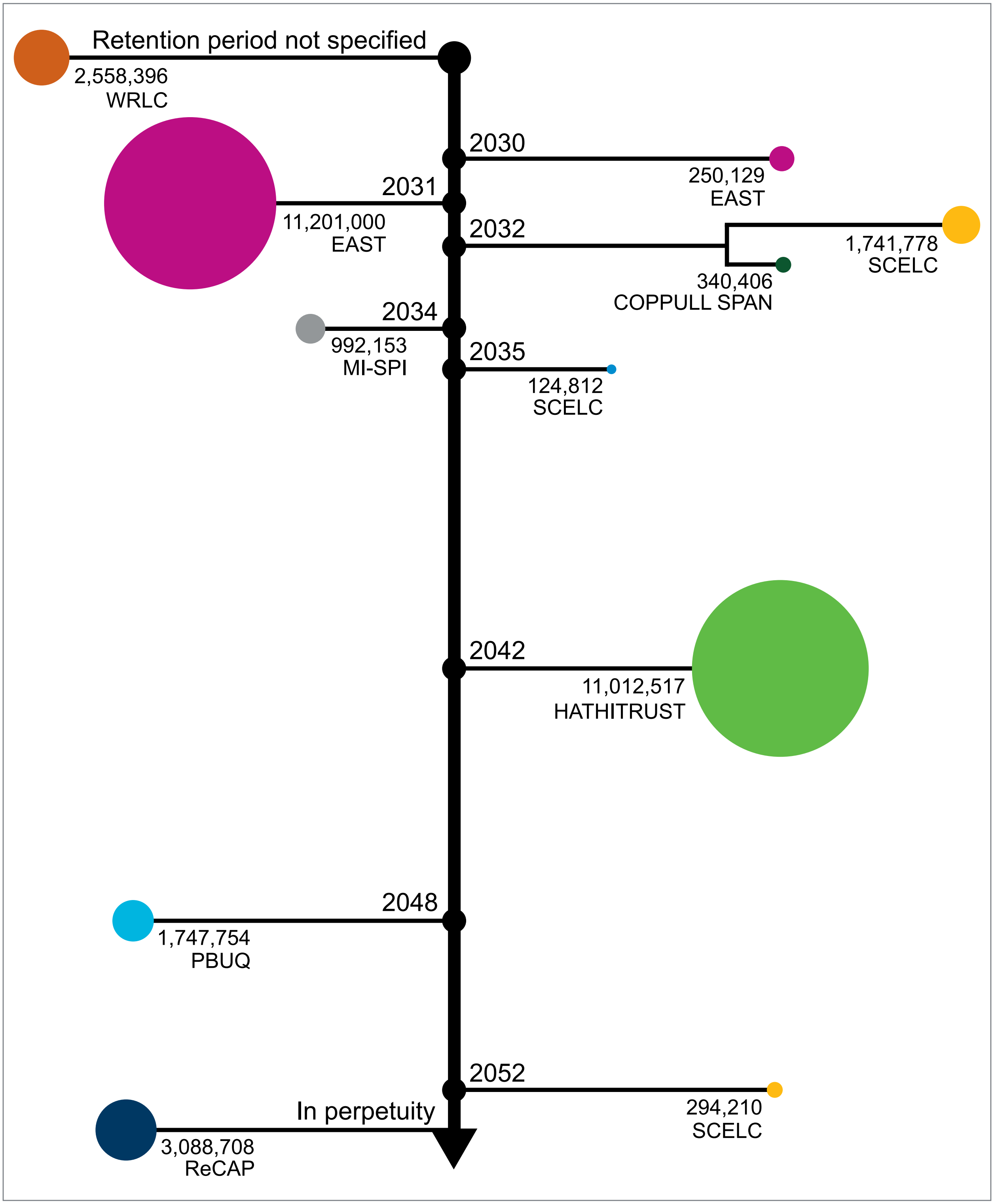


Figure 7. Timeline of commitment expiration by program

EAST has more than 11 million commitments set to expire by 2031, and HathiTrust has 11 million set to expire in 2042. While other shared print programs overlap some of those commitments, EAST and HathiTrust are so numerous in comparison that there will be titles that lose coverage if those commitments are not renewed. Some 6.1 million titles are at risk because they currently have expiration dates that are either already past or are scheduled to expire between 2028 and 2048. One smaller program, run by the ConnectNY consortium, had 218,000 titles whose retention commitments lapsed in 2023. Currently, there is no agreed-upon process for programs to work collaboratively to ensure commitment coverage on titles, either by moving copies—and the commitments—to an institution that can manage them, or finding other copies at institutions that are able and willing to make the retention commitment.

Only one program, the ReCAP service, has made a large number of commitments in perpetuity. Other programs, like the Florida Academic Repository (FLARE), Federal Depository Library Program (FDLP), and Scholars Trust, have registered a small number of perpetual commitments. These programs typically include a shared repository that, for storage of the retained titles, provides at least some level of validation for the presence and condition of the titles and provides access services, at least for the participating libraries. ReCAP's registered collection is 3 million commitments—just 8% of all registered commitments. Other programs have made it explicit that no retention period is specified. Washington Research Library Consortium (WRLC) has 2.5 million commitments (7%), which indicates that there is no guaranteed retention period. There are almost as many commitments without a guaranteed retention period as there are with a guaranteed perpetual commitment.

One other program, the Maine Shared Collections Cooperative, is currently reviewing its 1.3 million retention commitments scheduled to expire in 2028. It is anticipated that some—but not necessarily all—will be recommitted to by most of the member libraries. Over the next 25 years, almost 29 million retention commitments are scheduled to expire. EAST and SCELC have commitments totaling nearly 13 million scheduled to expire in 2031 and 2032, and both HathiTrust and PBUQ in Canada have expiration dates in the 2040s for another 12.8 million commitments.

As the pressures on academic and research libraries for space and budget continue to grow, even those programs committed to renewing their retention commitments for another 15–25 years (the most common retention periods) are likely to face resistance. The EAST program is already discussing ways to reduce the burden of retention on some of its earliest member libraries where the percentage of the collection already committed is highest. Community-based tools and infrastructure to support both program staff and individual libraries are needed to support this work across shared print programs. Further research to identify unique titles most at risk, as well as cross-program agreements to transfer commitments when needed, would assist in setting a clearer path for increasing the likelihood that the titles with commitments will still be protected and accessible for 25 years and beyond.

Community priorities: Commitment registration

This research aims to start discussions about strategies for shared print by identifying areas with inadequate or missing retention commitments. However, the findings are only as good as the available data, and there is good reason to believe that the data available to us is lacking. The benefits of submitting retention commitments to publicly available services have not been well-articulated and have received diminished attention over time. Furthermore, some programs and libraries are still figuring out how and when to make their commitments publicly available. For example, some of ReCAP's libraries are still working out internal processes to keep retention commitments up to date. This is no small feat because the program is actively adding hundreds of thousands of retention commitments annually.[9] There is no reason to think that ReCAP is alone.

Retention commitments also need to be adequate—that is, fit for purpose. The primary purpose of a registered commitment is to communicate to other institutions that an item is being retained along with details of time and manner. A corollary to that primary purpose is to document the commitment in a way that is understandable by machine processes. This requires the data elements to be expressed in a manner that is consistent across all registering institutions. Expressing the same information in unique or idiosyncratic ways undermines the purpose of the commitment by creating impediments to processing, thereby delaying or preventing the aggregation and dissemination of the information contained therein.

> **The shared print community is now poised to address areas that need more attention.**

Gaining an accurate picture of the shared print landscape is essential to deriving the greatest value from the collective effort. The shared print community is now poised to address areas that need more attention. Our data is good, but we are hopeful it will become more comprehensive and fit for purpose over the coming years as more libraries submit their commitments to publicly available databases. In the meantime, we hope that the shared print community can engage with this research and identify directions for further research and actions to take.

Calls to action

To address the risks and community priorities highlighted in this report, we make the following calls to action:

Keep the data coming

- Register all retention commitments—or as many as possible
- Create access to retention data to support the expansion and refinement of our print preservation strategies

Keep improving the data

- Invest in the improvement of metadata quality, which is essential for at-scale analysis and decision-making (particularly for identifying uniqueness)

Broaden the field

- Expand on this research to identify under-committed content areas
- Build more flexibility to allow for unspecified retentions (i.e., special collections) to be understood as a part of this landscape
- Bring more individuals and groups into the assessment and discussion of retention data

A commonly used set of tools or platforms to report commitments and analyze coverage quickly and easily would support these calls to action. It is necessary to convene the community to get their feedback on this research, synthesize their needs, and refine workflows to make reporting commitments easier.

CONCLUSION

Shared print is the effort to work together to meet an acceptable risk threshold by retaining an adequate number of copies. To meet that risk threshold sustainably, we need ongoing, more reliable, and granular data analysis to support informed print strategies and decision-making. Shared print is not intended to be a last-copy strategy—the approach many libraries employed a few decades ago.

It is unrealistic to think that every library has the means to commit to retaining copies in perpetuity. The shared print community needs to consider what to do when there is a scarcity of not only the number of titles, but also time, human capacity, and financial resources needed to meet a minimum threshold of risk tolerance even if all copies were committed. We do not know how often unique or scarcely held copies are being withdrawn each year. However, one can easily imagine technical infrastructure in the not-too-distant future that facilitates the migration of scarcely held but intellectually valuable items from smaller libraries that can no longer retain them to larger libraries with that capacity. There is reason to believe that many scarcely held materials could be outside the scope of shared print, residing in the many special collections, archives, or locally produced ephemeral collections.[10] Narrowing down the titles to those most suitable for shared print activities is a necessary step. Shared print alone will not preserve our collections. It is likely that even after reducing the scope of shared print from all titles in WorldCat to a narrower band, many titles will still be scarcely held.

Current automated cataloging workflows were created based on operational logic that is no longer congruent with today’s collective operational practices. Historically, knowledge organization descriptive practice centered on creating a local context for local use. As libraries increasingly evaluate local collections against collective collecting, it is imperative that metadata workflows are reexamined with an outward mindset. Recontextualizing local workflows within larger collective business needs, especially as they relate to data quality, is the first step to creating data collections that are fit for the purpose of collective collection stewardship. Investing in this change to operational practice will maximize individual organizations’ investment into metadata to gain the most collective benefit. As evidenced by the data engineering and analysis required for this research, improved metadata workflows and quality assessment, based on shared retention business operational needs, are fundamental to successfully moving the collective effort forward.

ACKNOWLEDGMENTS

The authors want to thank the Partnership for Shared Book Collections organization, specifically the Research and Network Level Standing Committee, for their interest and insight in forming this research endeavor, particularly Ian Bogus, Alison Wohlers, and former committee member Susan Stearns. We also want to recognize the collective effort of the OCLC Data Science and Metadata Research team for the data analysis and visualizations that made the narrative possible: Heather Carter, Gibi George, and Kirk Hess.

Many OCLC colleagues also supported this project by sharing their expertise on various aspects of shared print and collective collection operations and reviewing drafts of this manuscript to strengthen the report, including Andy Breeding, Heather Carlisle, Andrew Hall, Brian Lavoie, Scott Livingston, Constance Malpas, Chip Nilges, Bemal Rajapatirana, Mary Sauer-Games, Jenny Toves, Joanna White, David Whitehair, and Gina Winkler.

Lastly, the authors are grateful to the marketing team whose guidance and expertise were essential to the creation of this report: Amy Dalrymple, Jeanette McNicol, Erica Melko, and JD Shipengrover.

NOTES

1. A binational federation of 17 monograph shared print programs across the US and Canada. On July 1, 2024, the Partnership for Shared Book Collections and Rosemont Shared Print Alliance merged to form the Shared Print Partnership, which encompasses both serial and monograph formats.

   See "Merging Shared Print Organizations." 2023. *The Shared Print Partnership* (blog), 17 November 2023. https://sharedprint.org/shared-print-merger.

2. Bogus, Ian, Candace Arai Yano, Shannon Zachary, Jacob Nadal, Mary Miller, Helen N. Levenson, Fern Brody, and Sara Amato. 2023. "A Model to Determine Optimal Numbers of Monograph Copies for Preservation in Shared Print Collections." *College & Research Libraries* 84 (5): 767. https://doi.org/10.5860/crl.84.5.767.

3. These numbers are accurate as of August 2024.

4. "Best Practices for Discovery and Disclosure of Shared Print Items." 2020. *The Shared Print Partnership* (blog), 13 March 2020. https://sharedprint.org/best-practices/disclosure/.

5. Amato, Sara, and Susan Stearns. 2018. "Documenting the Stewardship of Libraries: The Eastern Academic Scholars' Trust Validation Sample Studies." *Collaborative Librarianship* 10(3), article 4, 168. https://digitalcommons.du.edu/cgi/viewcontent.cgi?article=1394&context=collaborativelibrarianship.

6. Bogus et al., "Monograph Copies for Preservation," (see n. 2.)

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

8. There are 51.5 million records from the 100.5 million US and Canada set that have an LC Class.

9. ReCAP believes there are an additional 7 million retention commitments unsubmitted as of the time of data analysis.

10. Schonfeld, "Books without Boundaries," (see n. 7.)