NASIG
TRANSFORMING THE INFORMATION COMMUNITY

Serials Analysis Directions: The importance of Standard Numbers in Collaborative Collection Development for Serials

Shannon Keller and Amy Wood

Abstract

The Research Collections and Preservation Consortium (ReCAP) partners contracted the Center for Research Libraries (CRL) to do a volume level analysis of their combined print serial collections. This paper focuses on the decision to use the OCLC number as the primary match point and the starting point for the collection analysis, challenges encountered due to inconsistent OCLC number use across collections, the way those challenges informed the overall project, and the method CRL created to find OCLC numbers for records without them. Final results of the volume-level analysis, ongoing work to integrate Harvard's serials into the analysis, and the effects of ReCAP partners' collaborative collection development initiatives are included.

Keywords: serials, collection analysis, consortia, collaborative collection development, shared collection development, Research Collections and Preservation Consortium (ReCAP), identifiers

Project Details

Background

The Research Collections and Preservation Consortium (ReCAP) started out between Columbia University, Princeton University, and the New York

https://doi.org/10.3998/nasig.4032

Public Library (NYPL). Harvard University has since joined the consortium in the midst of the grant project detailed within this paper. While Harvard participated in the project meetings, it was not officially part of the grant work, so its data were not evaluated as part of the original grant project.

The ReCAP partnership consists of a shared high-density storage facility on Princeton

University's Forrestal Campus, resource sharing services, and a shared collection development program. Resource sharing is managed through the individual library catalogs connected by a middleware software called Shared Collection Service Bus (SCSB), which was developed with funding from the Andrew W. Mellon Foundation. In addition to facilitating resource sharing, SCSB became the foundation for collection analysis tools that enabled ReCAP members to track usage, manage inventory in their shared facility and provide shared collection statistics to support collaborative collection development.

Collaborative collection development was a natural next step in the relationship between the ReCAP partners. With a successful and highly utilized mechanism in place to share resources housed at the ReCAP facility, the partners are embarking on shared collection development initiatives to expand the resources available to partner researchers. Beginning with certain foreign language materials, the partners are collaborating to determine the distribution of collecting responsibility of resource genres and subjects. Doing so will reduce the duplication of purchases, expand the number of items the partners are able to purchase and optimize acquisition budgets, both in terms of materials and space. Collaborative collection development of print serials has particular potential for collection cost savings, because serial runs take up significantly more storage space than individual monographs. Collaborative collection development of serials presents an opportunity to take advantage of all the mentioned benefits for significant long-term impacts.

In 2017, the ReCAP partners supported and participated in a project conducted by the Center for Research Libraries (CRL) to develop a method for identifying a master list of humanities and social sciences journals (Critical Corpus) for long-term preservation. The Critical Corpus project aggregated the print serial records from eighteen research libraries, including ReCAP partners Columbia University, NYPL, and Princeton University.[1] This project became the impetus for the ReCAP and CRL collaboration described herein. The Critical Corpus project resulted in a measure of bibliographic overlap among the three partners. It gave some insight into potential problems—namely, lack of numerical identifiers—that would prevent full sharing of collections and efficient, cost-effective stewardship of print serial holdings.

ReCAP leadership reviewed the project analysis and determined that ReCAP needed a method to better understand the duplication of serials at the volume or item level. With this knowledge, ReCAP leadership could better manage the ReCAP facility and future collaborative collection development efforts. The serials item level analysis project was born.

Project Goals

The serials item level analysis project was part of a larger Mellon-funded grant to expand upon the earlier development of SCSB, resource discovery, and collaborative collection development efforts. From the outset, the project's overarching goals were to identify holdings duplication, completeness, gaps, and uniqueness across the partners' serials collections in order to manage the combined collection from acquisition, through description, use and storage. ReCAP leadership could use information about collection uniqueness and duplication to develop retention, use and gap-filling rules and to plan space allocation.

Project Challenges

OCLC: Standard Number Solution

The heart of CRL's work was on the bibliographic reclamation and volume-level collection analysis. As identified in the critical corpus project, a bibliographic match point needed to be established at the title level to ensure we were comparing the same holdings. This is where standard numbers become extremely important, and we chose the OCLC number. Common naming conventions among serials, such as proceedings, index, newsletter, and review are rampant. Title changes are common and also make title-level matching difficult for serials. These are two of the reasons using a standard number is necessary for this process. Standard numbers facilitate the work to then normalize the holdings and analyze the level of duplication and uniqueness of holdings for the volume level analysis.

Why did the partners choose OCLC numbers as the solution for the bibliographic reclamation phase? There were several different options available to use when determining the best match point for the bibliographic reclamation phase of the project. As previously discussed, we decided to not use a text-based option such as title and focus on the standard number option because of the variance in title spellings and mistakes. There were really only two standard numbers that we seriously considered for the bibliographic match point: the OCLC number or the International Standard Serial Number (ISSN) number. The OCLC number was selected for two reasons. Firstly, the OCLC number was already the standard number used as a match point in SCSB. Secondly, because the ISSN number was not established until the seventies, it was much more likely that there would be an OCLC number for ceased or dead serials that predated the ISSN than there would be an ISSN. The likelihood of there being a bibliographic record for some of these older titles is very high, whereas it may not yet have an ISSN number assigned. During the first planning meeting, the project members readily decided to use OCLC numbers as the standard number for title-level matching.

Barriers and Challenges

Three major challenges the project team faced, included defining the project scope, managing a large group of stakeholders and addressing the results of cataloging practices that varied across partner institutions. The grant funded the hiring of a project manager, and Shelley Dexter was brought on to manage all aspects of the grant, not just the serials analysis portion. Shelley Dexter is a project manager by training and not a serials librarian, so she was helpful in making sure we clearly defined our project goals and scope.

Early on we decided to exclude newspapers, monographic series, and microforms from the project scope. The partners had widely varying catalog practices and policies for these formats, and it was decided that these materials were well outside the scope of the more booklike serials. At first, we thought we were only going to analyze holdings from the ReCAP facility, but eventually we decided to include all partner serials, even material stored on-site at their libraries, to be able to answer the question around completeness and gaps.

Managing a large group of stakeholders took finesse, but previous collaborative projects between the partners (including the development of SCSB), paved the way for successful collaboration. Sometimes different priorities between the partners and resources influenced project decision-making, but as all are committed and invested in the ReCAP partnership, this did not impact the overall success of the project. Mainly this manifested in the difficulty of finding a time when all stakeholders were available to meet.

From a technical, practical, project implementation perspective, the two biggest challenges were that NYPL had 87,000 records lacking an OCLC number and the partners had to decide, and agree, on the reporting format for the analysis. Overall, there were 96,000 records between the three partners lacking an OCLC number, but with such a large number coming from NYPL, a mechanism for identifying and validating OCLC numbers for those records was imperative. For comparison, there were approximately 92,000 records without an ISSN. But we suspected that it would be easier to find an authoritative OCLC number whereas the ISSN number could be included across multiple different bibliographic records. ISSN have been used only since the early 1970s and many serial titles held by the libraries pre-dated ISSN and would not have an ISSN.

Ultimately, the solution to produce reports at volume-level analysis involved developing a database. We initially considered using Excel spreadsheets but realized this would only result in a snapshot analysis of one point in time. By creating a database, wherein partners could search by title for examples, it is nimbler than Excel and allows for updates. This solution gave the partners flexibility to provide data updates, and for the ongoing analysis, including future integration of Harvard's serials data. To reach these goals, the project manager established four functional work phases: a bibliographic reclamation phase, the volume or item level collection analysis, building a report mechanism, and ongoing collection development.

Bibliographic Reclamation

Initial Assessment

The ReCAP libraries provided CRL with the full corpus of print serial records from their library catalogs for CRL to produce an initial assessment of the aggregated collection of print serials. This initial assessment was intended to be a lightweight overview to provide the partners and CRL with a common understanding of our starting point and to surface obvious errors in pulling the data from the catalogs or in CRL's approach to process the records. After reviewing the initial assessment, the project team decided to limit the scope of serials to include only periodicals and journals.

CRL began the initial assessment by validating all records against OCLC's WorldCat database. All ReCAP libraries were OCLC members. CRL extracted the presumed OCLC# from each record and deployed the OCLC search API From the API results, the team verified that the number we pulled from the record was an OCLC#, associated with the title, and whether it had been superseded. Bibliographic level and material type were also verified to exclude all records that were not describing print serials.

With the resulting record set of in-scope records, the CRL team provided key characteristics of the aggregated collection including coverage by country of publication, language and year of publication, and LC classification. A title-level assessment of overlap and uniqueness across the collections was included. The number of records without an OCLC number was also highlighted for discussion.

This gave us the foundation to build a path forward and determined a more nuanced scope for serials. At this point, the ReCAP team decided to exclude monographic series and newspapers. Both teams agreed that although a bibliographic reclamation for both formats would be beneficial, there were concerns that there would not be enough resources for this. In response, a set of test records was pulled from the full corpus to test a reclamation workflow and outline the phases of the reclamation and assessment for the overall project. The lightweight assessment laid the foundation for the project.

Strategy

The combined number of records without OCLC numbers totaled just over 95,000. Of those, 87,000 were for the NYPL. With that number of records, we needed a search strategy that allowed us to search OCLC's WorldCat database in batch and check batch results. There was no way we could search each of the titles one by one and no way that the ReCAP partners could review the results of the search one record at a time.

Using OCLC's Connexion Client, we developed a method that was scalable, transparent, and easy for any cataloger to emulate.[2] The process could be automated with APIs and word processing scripting languages like Python, but Connexion gave us the control we needed for this project.

Process

Our reclamation process had two steps: 1) querying OCLC's WorldCat database and 2) determining a good match.

Step 1: Querying OCLC's WorldCat database

Our reclamation process was iterative with a series of searches in batches of 5,000 records that started with the narrowest search possible and ended with broad keyword searches. Many records had identifiers other than OCLC numbers, therefore we started the searching with records that included both an ISSN and an LCCN. First, we deployed a search for records with both identifiers, then we switched to a single identifier. Once all records with identifiers were exhausted, we switched to keyword searching. Each type of search, whether identifier or keyword, was performed with a series of limits. We added OCLC symbols to the search to find a record that already had one of the ReCAP libraries' holdings attached.

Multi-search-term queries were also handled in batch. Using MARCedit, the following MARC fields and subfields were pulled from the MARC records supplied by each partner: fixed field (008), bytes 07–014, 15–17, 23, 35–37, and variable fields/subfields 110$a, 710$a, 245$a, 245$b, 245$p, 260$a, 260$b, and 362$a. Content from those fields and subfields were paired with corresponding WorldCat index labels to create batches of complex derived searches. A typical first search for a record would look like: li:nyp mt:cnr mf:nmc ll:eng yr:1884 au=Verein für die Geschichte Berlins pl:Berlin ti:Zeitschrift. A colon or equal sign was used between the index label and the search terms depending on whether the search was keyword or phrase respectively.

Search strings were custom made for each record containing the data pulled from the partner records, and each pass used the same search criteria for each record until all records were searched. Results were checked and separated into successful and unsuccessful results. Records with unsuccessful results were searched again with one fewer search index and term, and in additional searches if necessary. Each subsequent search removed a search index and term. For more information on derived searching and WorldCat indexes and corresponding labels, see OCLC's Connexion: Searching WorldCat quick reference.[3]

The process of searching with the most specific terms, like an identifier, using limits like OCLC symbol, before moving to broad keyword searches was designed to return the fewest matching records as possible to determine a good match.

Step 2: Determining a good match

With 96,000 records to search, reviewing the search results was critical to ensure that the correct record was found. With input from the CRL reclamation team and ReCAP partner catalogers, a simple system that compared six fields in the local record contributed by the partner to the corresponding field in the retrieved WorldCat record, gave each match a number 1 and each non-match a 0 and totaled the sum of matches. When all fields matched, the found record was given a score of six. When any five fields matched, the record was scored a five, and so on. Records with scores of five or six were considered acceptable. Records with lower scores were examined more closely, and sometimes received a pass, but most often were marked for additional searching in WorldCat. Fields used included: 245, 110, 710, 008 byte 07–10 (date1), 008 byte 15–17 (country), and 008 byte 35–37 (language).

Using this method also allowed us to question whether a match was considered the best record. There were several titles with duplicate records in WorldCat, and for various reasons, the partners did not always have the best record. Perhaps another library contributed a record to WorldCat after the partner's record was cataloged and the two were never merged using OCLC's regular deduplication process. During the project extension, CRL's team identified records with less than full level cataloging and did additional searches for records with more complete cataloging description and access. Consideration for partner integration of identified OCLC numbers in local holdings is ongoing.

Results

The bibliographic reclamation included 96,200 records. Of those 91 percent (87,904 records) were reconciled. The additional nine percent had too little information to create a worthwhile search or prevented us from using our checking system outlined in Step 2 above. About 23,000 of the 87,904 were out of scope as monographic series, newspaper, or non-print format, but lacked the coding in the local record to enable us to exclude them after the initial assessment.

Summary and Key Takeaways

At the beginning of the project, there was much discussion about collection management and the idea that understanding the level of duplication or completeness at the volume level would give the partners the ability to perhaps deduplicate some of the holdings at the ReCAP facility. The recognition to save some space and then utilize that saved space for future collecting is an ongoing need between the partners. In reality, this project identified that there is not significant duplication across the serials holdings, even between three large research institutions such as Columbia, Princeton, and the New York Public Library, at least not such that would warrant a deduplication effort at this time. This project highlighted the opportunity for the ongoing collaborative collection development and prospective deduplication of serials by expanding the number of serials the partners can collect by only one partner collecting that title and making it available in the shared collection. As previously noted, this will allow the partners to collect a more resources across many more languages and disciplines and collect more deeply than they have been able in the recent past.

Harvard's influence on the project should not go unacknowledged. When determining how to report out the results of the volume-level analysis, we realized that doing static reports would not provide us the opportunity to update the data or later integrate Harvard's serials into the analysis. Choosing to create a database format for the volume-level analysis report gave us a lot of room to grow, do future analyses, and integrate other institutions' data, which the team at CRL and the ReCAP partners are doing with Harvard's data.

The partners gained great value in understanding one's collection. Moving forward, the partners can focus on collaborative collection development projects knowing it is the best path forward for the consortia.

Acknowledgements

The authors acknowledge that this project was generously funded through the Andrew W. Mellon Foundation. Thanks to the partner libraries Serials Working Group and the staff at CRL working on this project.

Partner Libraries Serials Working Group:

- Abdul Alsaidi, Manager Cataloging, NYPL
- Jennifer Baxmeyer, Leader Western Languages Cataloging Team, Princeton University
- Joyce Bell, Cataloging and Metadata Services Director, Princeton University
- Christopher Cronin, Associate University Librarian for Collections, Columbia University
- Shelley Dexter, Project Manager, ReCAP Shared Collections
- Anthony Fischetti, Collections and Services Analyst, NYPL
- Denise Hibay, Astor Director for Collections and Research Services, NYPL

- Shannon Keller, Helen Bernstein Librarian for Periodicals and Journals, NYPL
- Heide Miklitz, Assistant Director for NYPL Research Materials, NYPL
- Steven Pisani, Assistant Director Cataloging, NYPL
- Robert Rendall, Principal Serials Cataloger, Columbia University
- Steven Riel, Manager of Serials Cataloging, Harvard University
- Lyudmila Shpileva, Serials Cataloger, NYPL
- Marie Wange-Connelly, Head of Circulation and Inventory Management Systems, Princeton University
- Mark Wilson, Director of Monographs Processing Services, Columbia University
- Breck Witte, Director of Library Information Technology, Columbia University
- Mark Zelesky, Integrated Library System Coordinator, Princeton University Center for Research Libraries:
- Amy Wood, Head of Technical Services, CRL
- Yoseline Louisma, Special Projects/Program Manager, CRL
- Nathaniel Florin, Library Specialist, CRL
- Andrew Elliott, Functional Specialist, CRL
- Stephen Early, Senior Cataloger, CRL
- Jenna Mosillami, Cataloging Assistant, CRL

Contributor Notes

Shannon Keller is the Electronic Resources Librarian with the U.S. Department of State's Ralph J. Bunche Library. At the time of the project detailed in this paper she was the Helen Bernstein Librarian for Periodicals and Journals with the New York Public Library.

Amy Wood is the Head of Metadata & Discovery Enhancement at the Center for Research Libraries.

Notes

[1] "Critical Corpus report summary," Center for Research Libraries, Global Resource Network, accessed July 20, 2022, http://www.crl.edu/sites/default/files/event_materials/Critical%20Corpus%20Analysis%20Status%20Report.pdf.

[2] "Connexion Client," OCLC, last modified July, 13, 2022, https://help.oclc.org/Metadata_Services/Connexion/Connexion_client.

[3] "OCLC Connexion: Searching WorldCat Quick Reference," OCLC, accessed Dec 11, 2022, https://help.oclc.org/Metadata_Services/Connexion/Connexion_client/Reference/Quick_reference_Searching_WorldCat_in_Connexion_client_and_browser?sl=en