IFLA WLIC 2019 Athens

Submitted on: 31.05.2019

Using standard numbers to conduct a serial item-level holdings analysis of the ReCAP partners' collections

Shannon Keller
Collections and Research Services, New York Public Library
New York, United States of America

Amy Wood
Technical Services, Center for Research Libraries
Chicago, United States of America

Abstract:

The Research Collections and Preservation Consortium (ReCAP) is one of the largest shared print, preservation, and collection repositories in North America. Through funding from the Andrew W. Mellon Foundation, ReCAP's founding Partners (Columbia University, New York Public Library, and Princeton University) are undertaking a cross-collection analysis to identify the level of duplication and uniqueness in the print serials across their collections in order to inform collection development and management decisions. ReCAP partnered with the Center for Research Libraries (CRL), to undertake a bibliographic reclamation and item-level holdings analysis. The project is creating methodologies and workflows to identify possible records in OCLC's WorldCat, to review results at scale, and to use the results to aid in a concurrent item-level holdings analysis. The analysis includes a review of roughly 660,000 bibliographic and nearly a million item or holdings records.

An initial review of the records revealed over 96,000 serial records lacking an OCLC number, and only 14% of the records without an OCLC# had either an ISSN, or an LCCN. The item-level holdings analysis utilizes normalized and actual holdings data to expose detailed information about ReCAP's serial holdings. After nine months of work, over 49,000 WorldCat records were found as possible matches for ReCAP records, and only 406 ReCAP records were determined to lack sufficient information to identify possible matches.

The results of this analysis will inform future serial subscription purchases, as well as de-duplication decisions, gap filling, and identification of full runs through combined collection holdings. Methodologies established in this project may be applied to the serial collection of ReCAP's newest partner, Harvard University, as they prepare to both send items to the facility and expand access to their collection through the partner's shared collection services.

Keywords: Serials, Shared Print, Collections Management, Consortia, ReCAP, Center for Research Libraries.

Introduction

The activities described in this paper were undertaken by the Research Collections and Preservation Consortium (ReCAP) and the Center for Research Libraries (CRL) as part of the ReCAP Phase III Discovery to Delivery: Strengthening Operations and Sharing grant funded by the Andrew W. Mellon Foundation to develop and implement a Shared Collection from among the three partners' ReCAP holdings. Phase one established the Shared Collection. Phase two developed a Shared Collection Service Bus (SCSB or "middleware"), which allows patrons from all three libraries to search for, request, and borrow materials from the Shared Collection, regardless of the item's library of origination. Phase three will integrate governance of the Shared Collection and its related technology into the ongoing operations of ReCAP; document the middleware to encourage adoption by other shared archives; and enhance the partners' ability to build and manage the ReCAP Shared Collection. The item-level serials comparison, tackled in the current phase, will be the foundation upon which partners normalize their data, analyze, build and manage the Shared Collection in the future.

Organizations

The Research Collections and Preservation Consortium (ReCAP) is one of the largest shared print, preservation, and collection repositories in North America. ReCAP's facility resides on Princeton University's Forrestal Campus in Princeton, New Jersey. Since 2012, the founding partners of ReCAP, Columbia University, New York Public Library (NYPL), and Princeton University, have developed and implemented a Shared Collection built from the three partners collections at ReCAP. The Shared Collection allows a patron from the partner institutions to borrow materials from across all three partner library collections seamlessly, regardless of the item's owning library origin. An integral component of the Shared Collection Service Bus (SCSB) is the title-level matching algorithm which utilizes titles and standard numerical identifiers to identify duplication across monographs and serials at the title level, and designate one copy as the 'shared' copy.

The Center for Research Libraries (CRL) is an international consortium of university, college and independent research libraries. Founded in 1949, CRL supports research and teaching in the humanities, sciences, and social sciences by preserving and making available a wealth of rare and uncommon primary source material and published resources from around the world. Since 2012, CRL has undertaken a variety of activities to support strategic, coordinated efforts to manage and preserve essential print serial collections. Key activities and tools to support shared print activities include: establishment of the Print Archive Network (PAN) Forum[i] to discuss best practices and share information, the Print Archives Preservation Registry (PAPR)[ii] to freely disclose holdings committed to shared print collections, and a variety of record validation, holdings normalization, and collection comparison activities.

Background of Collaboration

In 2017, ReCAP libraries supported a CRL project, entitled Critical Corpus, to plan and measure the strategic print preservation efforts of North American libraries for Social Sciences and Humanities[iii]. The goals of the Critical Corpus project were: to define the costs and requirements for preserving the "universe" of Humanities and Social Sciences serials; to develop and cost out a methodology and strategy to identify the "critical corpus" of journal literature published in print form and important to academic research in the humanities and social sciences; and to develop a significantly large list of titles to lay the groundwork for review and curation of a final list. The project aggregated the print serial records from eighteen research libraries, including ReCAP partners Columbia University, NYPL, and Princeton University. This project became the impetus for the current ReCAP and CRL collaboration, described herein. The critical corpus project provided a measure of bibliographic overlap among the three partners and gave some insight into potential problems—namely lack of numerical identifiers—that would prevent full sharing of collections, and efficient, cost-effective management of and access to holdings.

After reviewing the results of the Critical Corpus project, ReCAP leadership proposed going deeper, to quantify overlap and uniqueness at the holdings level. All libraries have missing, incomplete or damaged issues and volumes within their serials collections; a comparison performed at the title level masks all of that important information. A title-level analysis is important, but it is only the first step in establishing true overlap.

Since 2012, CRL has made various attempts to create issue level comparisons between record sets. Even with granular issue level data like the data produced or managed by responsible repositories of digital serials such as JSTOR, CLOCKSS, and Portico, and by responsible providers of digital versions of print serials, the task of mapping library holdings to issue level data remained a problem without a solution that would scale up beyond a small, discrete set of titles with predictable publication patterns.

Funding from the Andrew W. Mellon Foundation for ReCAP's third phase of Shared Collection project development provided the opportunity and resources to apply what CRL learned in earlier attempts at issue-level comparison toward developing a solution that would enable ReCAP to analyze and compare their collections at the item level. In doing so, CRL and ReCAP would develop a collection comparison model that would be reproducible for future phases or further collection partners. This paper will outline the planning, processes, and decision making workflows utilized to conduct this analysis.

Statement of the problem

As part of the Shared Collection, ReCAP partners committed to managing materials in accordance with agreed policies, including but not limited to: retention in perpetuity and usage rules (borrowing, in-library use, and supervised use). To best manage the Shared Collection, the partners endeavored to understand duplication, completeness, and gaps in their serial holdings. Developing methods for understanding these facets of a serial title run requires the ability to compare partner holdings at the volume or item level.

The item-level holdings analysis of the ReCAP partner's serial collections funded through this grant will streamline partner efforts to efficiently manage storage at the facility, identify titles wherein partners retain the complete run of a title, manage future serial transfers to the facility, and explore future options for shared collection development of serials.

Process and workflow

Planning

The grant phase began in January 2018 and will end in December 2019. Hiring began in January and three planning meetings were held to organize staff, confirm goals, develop procedures and establish a timeline of activities and deadlines within the project.

The first meeting, held in April 2018, was an in-person meeting at New York Public Library with all partner and CRL representatives attending. The initial meeting was focused on planning, identifying lead contacts for partner institutions, confirming what records should be shared with CRL, and the process by which those records would be shared.

A second in-person meeting was held in July 2018 after CRL had aggregated the records of the partners and validated important elements in the bibliographic records against corresponding records in OCLC's WorldCat database. The aggregation of records established the size of the full data set CRL would be processing and an estimate of overlap among the partners' print serial records. The validation, described in the next section, ensured increased accuracy for establishing and categorizing the records in the data set. Ideas for tools to share data between the CRL team, who are based in Chicago, and the ReCAP team, who are based in New York City, were proposed.

A third meeting further refined the goals, timeline, and ways of sharing data. One of the most important things coming out of this meeting was a tool to assign level of confidence to the results of the bibliographic reclamation. The bibliographic reclamation reviewed local records that lacked an OCLC number. The goal of the reclamation was to find a matching record in OCLC's WorldCat database using available information in the local record to search for a match.

Project Scope

The project was defined in two parts: a bibliographic record reclamation phase, and an item-level holdings analysis phase. Integral to the item-level holdings analysis is the bibliographic record reclamation to identify standard numbers for partner matching at the title level. The partners choose to include in the analysis their entire serial bibliographic record set, not just the records for titles at ReCAP. This was decided in part because the partners are regularly trying to make decisions about which titles and items to send to ReCAP from their on-site storage facilities, and a better understanding of the duplication or uniqueness of a serial title would help inform these decisions. It will also inform future collaborative collection development efforts. In addition, during the planning phase partners identified areas that are out of scope for the item-level holdings analysis, including: microfilm, monographic series, book sets, and newspapers.

Bibliographic record reclamation

In order to meet the goals of the project to measure overlap and uniqueness among ReCAP partners, it was essential to be able to include all of their records in the cross-collection comparison. From past experience using bibliographic records to execute comparisons of U.S. library collections, CRL has found most success using OCLC numbers as the primary match point. The pitfalls of using OCLC numbers as a match point are well known, and could be factored into the work. The work was carried out in distinct phases and was expected to take eight months. Difficulties finding matches, and balancing the work of the project with other priorities stretched the deadline to twelve months.

Starting Point

Based on what was learned about the ReCAP partners' bibliographic records from a previous project, CRL knew that the bibliographic records would reflect a high level of investment in bibliographic description, but they would also reflect changes in cataloging rules over time, varying local practices, and levels of cataloging due to staffing expertise. The first step for this project was to quantify and document those differences and discern whether there would be particular challenges that needed to be addressed to reach the project goals.

Each of the ReCAP partners provided a file containing all print serial records. The combined record set included approximately 660,000 records. NYPL records made up 55% of the record set and Princeton and Columbia had 24% and 21% respectively. Within each institution's set of records, CRL identified fields most likely containing OCLC numbers. The MARC 003 field should indicate which field has the OCLC number, but it was frequently wrong. The 001, 035, and 079 fields were the primary sources of OCLC numbers, but a small number of records were found in other fields. The prefixes OCoLC, ocn and ocm were sought, but some records did not include these with their OCLC numbers. Because the 001 and 035 are also used for other identifiers besides OCLC numbers, and those identifiers can be mistaken for OCLC numbers, CRL checked all suspected OCLC numbers in a subsequent validation step.

Validation

Partners' records were validated against OCLC's WorldCat database. Although not without error, records in WorldCat, as an internationally shared repository of bibliographic information, should reflect a closer alignment to cataloging standards than a local catalog where records may be edited for practical reasons to suit the perceived needs of a library's patrons.

The OCLC numbers gleaned from the ReCAP partners' records were used to initiate an API call against the WorldCat database to pull corresponding bibliographic records. Key fields were then pulled from those records for validation. Fields included: OCLC number, superseded OCLC number, ISSN, title, imprint, fixed field date 1, fixed field date 2, country code, language code, bib level, material type, and serial type.

CRL reported differences between the local records and the WorldCat records to the ReCAP partners. ReCAP partners concluded that reviewing reported discrepancies and identifying correct information was beyond the scope of the project. Only bib level, serial type, and material type would be used to weed out-of-scope records─that is, anything other than print journals. Title matching would be used to call out identifiers mistakenly extracted as OCLC numbers. If a number, thought to be an OCLC number, retrieved a bibliographic record with a title clearly different than the local record, that number was not accepted as the OCLC number for that record.

Aggregation and Overlap

The record review and validation resulted in a combined record set of approximately 453,300 unique titles and about 96,400 records without OCLC numbers. Nineteen percent of the unique titles, roughly 87,800, were held by more than one partner. Four percent were held by all three partners and 15% were held by two partners. Eighty-one percent were held by a single library.

The record review and validation resulted in the identification of 95,778 in-scope records without OCLC numbers.

Finding OCLC number for bibliographic records without them

Records without OCLC numbers were divided into four groups for each partner: those with ISSN and LCCN, those with ISSN, those with LCCN and those with no numerical identifiers. Number of records for each category were:

Categories	NYPL	Columbia	Princeton	Total Records
ISSN & LCCN, no OCLC	8,795	19	166	8,980
ISSN no OCLC	2,836	12	901	3,749
LCCN no OCLC	15,921	49	185	16,155
No ISSN, LCCN nor OCLC	59,895	683	6,316	66,894
Total Records with no OCLC	87,447	763	7,568	95,778

Prior to the ReCAP project, CRL had extensive experience using OCLC's Connexion client cataloging software—a simple yet powerful tool—to search OCLC's WorldCat database, download, edit and organize large record sets for other cataloging projects. Connexion client's batch searching was the essential feature needed to process over 95,000 records within the grant period.

In preparation for batch searching in the WorldCat database, essential fields of data were pulled from the local records, including: institution, OCLC number, ISSN, LCCN, title, author, publisher, place of publication, date 1, date 2, country code, and language code. This field data was then stored in an MS Access database until needed for the searches.

Note: in this project, cataloging class descriptors were not viable search fields. NYPL uses a native fixed order classification scheme which limits the ability to us common classification schemes such as the Library of Congress System.

Batch searches are performed with search keys and text in the search syntax of a command line search catalogers use to search for individual records in an interactive session. OCLC offers documentation Connexion client documentation, including batch search instructions, on the OCLC webpage[iv]: https://help.oclc.org/Metadata_Services/Connexion/Connexion_client_documentation

All batch searching was an iterative process starting with as many search keys for which the local records had information. OCLC library symbol, material type, ISSN, LCCN, where available, language, country and place of publication, dates of publication, author, and title were all used in various searches. The goal was to retrieve as few records as possible that matched our search criteria. Even specific search terms like an ISSN or LCCN can retrieve many, many records. In a few extreme cases, even using the library symbol to limit results would reveal that a partner library had their holdings attached to more than one record for a single title. If searches failed to retrieve records, single search keys and their associated terms were removed and the revised search string was included in a subsequent batch.

Records that contained an ISSN or LCCN rarely, if ever, required the addition of keywords from title, author or place of publication. However, limiting the numeric searches with language, country of publication and dates of publication were crucial to finding the correct record since ISSNs are often applied incorrectly to a bibliographic record.

Records with no numerical identifier required a keyword strategy. Words from title, author, and publication location replaced the identifier for searching. A Python script was created to normalize the text strings by taking out diacritics, articles, punctuations, some common words like: “for”, “and”, “some”, “et”, “und”, “etc”, “or”, “on”, etc.

Sample search strings include:

Search strings for records with ISSN, LCCN or both
li:nyp ll:eng mt:cnr mf:nmc in:1234-5678 yr:1962 pl:paris
li:nyp ll:eng mt:cnr mf:nmc ln:5678991 yr:1962 pl:paris
li:nyp ll:eng mt:cnr mf:nmc in:1234-5678 ln:9587823 pl:paris yr:1962

Search strings for records without ISSN or LCCN
li:nyp ll:eng mt:cnr mf:nmc yr:1962 pl:paris ti:dictionnaire basque-francais
li:nyp ll:eng mt:cnr mf:nmc yr:1944 au:universidad de panama pl:panama ti:boletin
li:nyp ll:eng mt:cnr mf:nmc yr:1944 au:church of jesus christ of latter-day saints pl:salt lake city ti:m i a dance handbook
li:nyp ll:eng mt:cnr mf:nmc yr:1921 au:great britain pl:london ti:report on economic conditions in algeria tunisia and tripolitania
li:nyp ll:eng mt:cnr mf:nmc yr: au: pl:sl ti:a collection of booksellers' and auctioneers' catalogues
li:nyp ll:eng mt:cnr mf:nmc yr:1830 pl:london ti:a penny paper for the people
li:nyp ll:eng mt:cnr mf:nmc pl:buenos aires ti:acta

Search key	**Meaning**
li	OCLC holding institution symbol
ll	language
mt	material type (cnr was used for “continuing resources”)
mf	microform (nmc was used for “not microform”)
yr	date 1 and date 2
in	ISSN
ln	LCCN
au	author
ti	ti
pl	place of publication

Batches were searched in groups of 5,000 search strings. To prepare the records for searching, data was pulled from the local MARC records and compiled in an MS Excel spreadsheet. Search strings were compiled using a simple Excel concatenate function with fields listed in the table below.

Essential MARC fields	Corresponding search key
BIB	local reference
date1	yr
date2	yr
Cntry	local reference
Form	local reference
lang	li
362$a	yr
normalized_110a	au
normalized_710a	au
normalized_245a	ti
normalized_260a	pl
normalized_260b	pb

A batch of 5,000 searches took about an hour or two to run. Initial review of the results would take several hours to complete. The time needed for review depended on the complexity of the search strategy and the number of resulting records.

Results from the WorldCat searches were exported as MARC records from Connexion Client. MarcEdit was used to extract OCLC number, LCCN, ISSN, date, title, publisher, and date of publication data from the MARC records. Data was then imported into MS Excel to remove duplicates and records that were obviously out of scope, such as hybrid electronic records cataloged as print with 856 link fields. Once the data was in a spreadsheet form, cataloging assistants could review the results manually line by line.

Successful results from all query strategies were compiled in a list. Six fields from the WorldCat records were compared to the corresponding fields from the partners' records to confirm a match. Records that did not retrieve records from WorldCat were marked to be included in more searches with slightly different strategies. When all possible batch search strategies are exhausted, remaining records are searched manually one by one.

Approval of OCLC numbers

The final step in the bibliographic reclamation is for ReCAP partner libraries to approve the OCLC number and corresponding record found in OCLC's WorldCat. In an initial sample, results from the queries returned some incorrect records for various reasons, primarily the generic characteristics of the search terms.

With almost 96,000 records to review, ReCAP partners needed a tool to help determine level of confidence in the results of the query. This would help them budget their resources where they were needed most. ReCAP and CRL agreed on six essential fields in the bibliographic record, and a simple algorithm for those fields to determine level of confidence.

Each of the six fields in the local record was matched against the corresponding fields in the presumed matching WorldCat record. Each field that matched was given a score of one and each field that did not match scored a zero. Results for each field were added. Those records scoring a six and a five, if there was a title match with the score of five, were deemed the highest level of confidence and were the first group of records for CRL to send to the ReCAP partners for review. Records that scored a lower number were put through additional review and searching by CRL.

CRL used an MS Access database with the parsed local records and parsed OCLC bibliographic records to perform the comparison and a spreadsheet on a shared drive to present the results to the ReCAP partners.

The Access query performing the comparison of the six essential fields is:

```
        SELECT [OCLC BIB].INST, LIB.INST, LIB.[LIB bib], [OCLC BIB].[WC oclc],
    LIB.[LIB title], [OCLC BIB].[WC title], LIB.[LIB 110], [OCLC BIB].[WC 110],
    LIB.[LIB 710], [OCLC BIB].[WC 710], LIB.[LIB date1], [OCLC BIB].[WC date1],
    LIB.[LIB cntry], [OCLC BIB].[WC cntry], LIB.[LIB lang], [OCLC BIB].[WC lang],
    IIf([LIB title]=[WC title],1,0) AS title, IIf([LIB 110]=[WC 110],1,0) AS 110, IIf([LIB
    710]=[WC 710],1,0) AS 710, IIf([LIB date1]=[WC date1],1,0) AS date1, IIf([LIB
    cntry]=[WC cntry],1,0) AS Place, IIf([LIB lang]=[WC lang],1,0) AS Lang,
    [title]+[110]+[710]+[date1]+[Place]+[Lang] AS Total, [OCLC BIB].Note
        FROM LIB INNER JOIN [OCLC BIB] ON LIB.[LIB bib] = [OCLC BIB].[bib
    no];
```

Further defining goals and deliverables

Following the initial phase of bibliographic record reclamation and with feedback from the partner libraries and in conversations with CRL, the overarching goals and deliverables of this grant phase were further defined.

Goals:

- Understand the scope of the print serials record remediation and identify workflows for record correction
- Gain a preliminary understanding of the approaches and methodologies of matching at the item-level for the item-level holdings analysis
- As much as partner bandwidth allows, perform serials records cleanup in order to facilitate the holdings analysis, an item-level comparison, of the partners' collection

These goals gave specification to the project as outlined in the original grant proposal. They recognize the human investment needed to improve serial bibliographic records, which will also help facilitate the item-level holdings analysis.

Holdings / Item record data normalization

Understanding true overlap of holdings among ReCAP partners required more than a bibliographic comparison. A bibliographic comparison of records identified approximately 19,600 titles held by all three partners and just over 68,000 held by two partners. However, title level comparisons often mask incomplete runs. An essential component of this project was to complete a comparison at the item level. The primary challenges for comparing holdings below the title level across library collections are the differences in expression of holdings and how holdings data is stored in the local catalogs.

It was important for CRL to retain the ReCAP partners' holdings expression as they were, but also to create a means of mapping the holdings to a common item in order to identify overlap or gaps in holdings. CRL and ReCAP partners agreed that working at the volume level was sufficient. Any holdings expression of less than a complete volume were mapped to its corresponding volume. Multiple items such as "v.3:no:1-6" and "v.3:no.7-12" would both be mapped to "v.3" in an MS Access database, where all data would be stored and delivered to the ReCAP partners at the end of the project.

Data in the bibliographic and holdings records was used in the attempt to compile a full list of volumes for a title. Data from beginning and end publication dates, MARC 3XX fields, item records and holdings records was considered for each title to create the list of volumes. Holdings that appeared to be before or after the publication dates were recorded with a volume 0 or volume z respectively. Holdings that could not clearly be mapped to a volume were assigned a volume @. The procedure for creating the canonical volumes and mapping the holdings to the volumes is listed below.

Creating Canonical Volumes and Matching To Observed Holdings

- Identify the set to be worked on – titles held by one institution, two institutions, or all three – based primarily on overlapping OCLC numbers.
- Fetch WorldCat bibliographic records for all titles in the set, using Python scripts to interact with the WorldCat API.
- Using WorldCat bib data, remove out-of-scope titles from the working set. Titles are considered out of scope if they are:
    - Not serials (e.g., monographs)
    - Monographic serials
    - Not hard copy (e.g., electronic, microform)
- Using Python MARC processing scripts, extract frequency data from the WorldCat MARC and export it to a file. The specific fields are:
    - 310 (current frequency) and 321 (former frequency)
    - Date 1 and date 2 from the 008
    - All 362 (publication date) fields
- Using local MARC processing scripts, extract holdings data from all ReCAP records in the set and print it to a spreadsheet.
    - Each individual line of holdings data is stored separately in the spreadsheet, so a record with 10 item records and one summary statement will have 11 lines in the spreadsheet.
    - The type of holdings line (summary statement or item record) is recorded. For this set, any holdings pulled from an 863 to 868 line is considered a summary statement, all others are considered an item record.
    - Supplement and index holdings are identified. Holdings found in 864 and 867 fields are considered supplements, holdings found in 865 and 868 fields are considered indexes.
- Using Python normalizing scripts, separately normalize each holdings line and record the results in an output spreadsheet.

    - Since we are working at the volume level and/or year level, detail below that level is often excluded. For example, “v.12 pt. 3 (Apr 12, 1992)” might become simply “v.12 (1992)”. This prevents false precision and sidesteps a lot of errors.
    - Non-English and non-standard terms are normalized to “volume” and “number”. Terms for series statements are normalized to “new series” or “series 1”, “series 2”, etc.
    - Supplements and indexes present in item records or in summary statements found in the 863 and 866 fields are extracted at this point and recorded with the other summary/index data. For example, from the item record “v.12 (1992); index 1989-1991" the segment “index 1989-1991" will be shunted off to the index output file.
- Using Python normalizing scripts that are substantially the same as the above, normalize the 362 data from the WorldCat MARC and save to a spreadsheet. Pay special attention to terms indicating start or end volumes and years.
- Using a set of local Python scripts, pull out any series, volume, and year data from every normalized holding line and record the output in a spreadsheet.
    - For titles where “number” is the primary level of enumeration (“no.1 (1992)-no.128 (2003)”), “number” is treated as “volume”.
    - Where there are large runs of years or volumes, nonsensical ranges (“v.20-v.6”), or other obvious issues, the operator will be asked to look at the original data and confirm, alter, or reject the normalized output.
- Using a set of local Python scripts, create canonical volumes for sets of titles with two or three holding libraries.
    - Identify the start and end volume/year pairs for the title. These are done by examining a combination of local holdings and data from the MARC 008 and 362 fields.
    - Identify intervening canonical volumes between the start and end pairs. When possible, base canonical volumes on volumes and years actually seen in the holdings, rather than projected from publication patterns.
    - For titles held by only one institution, canonical volumes won’t be created. Instead only the start and end date for the title, based on WorldCat MARC, will be recorded.
    - Canonical volumes are printed to a spreadsheet so they can be added to the database.
- Using local Python scripts, match actual holdings to the canonical volumes and upload them to the database.
    - Holdings lines are linked to every canonical volume that they match. So the holdings line “V.8-v.12” could be matched to five different canonical volumes.
    - Holdings that come before or after the start and end of the canonical volumes get special “too early”/”too late” codes in the database.
    - Holdings that can’t be interpreted or that don’t include actual holdings data (“Check at circulation desk”) get a special “can’t interpret” code in the database.
    - Holdings lines that don’t exactly match a canonical volume will be linked to the best guess, or given the “can’t interpret” code if they can’t be reasonably attached anywhere.
    - Print the output to a spreadsheet to be uploaded to the database.

Item-level holdings analysis

While CRL created the process for normalizing the item level holdings data and comparing volumes across partner libraries, the partners further discussed the potential use cases for the item-level holdings analysis in order to help determine the best possible format for the deliverable analysis from CRL. From the outset, the partners expected to use the item-level holdings analysis to conduct de-duplication of the partners' serials collections in the Shared Collection. In addition, the partners anticipated using the analysis to identify complete runs of a title through their shared holdings, to help manage ongoing print serial subscriptions, and to identify gaps in a title.

Once CRL began normalizing and aggregating the item-level data, it quickly became apparent that MS Excel could not accommodate the item-level analysis, as the combined collection was too large for results to be stored, viewed, and manipulated easily as spreadsheets. The partners and CRL discussed potential options and decided on an MS Access database as the best format in which to deliver the results of the item processing. With feedback from the ReCAP partners, CRL created the database and a core set of queries and reports. Some examples include:

- Number of titles held by each partner
- Number of pieces expected per title and the number of pieces held by each partner per title
- Missing items
- Incomplete items (based on notes in the partners' holdings)
- Lacking items by title

The partners are still determining who will maintain the database at the conclusion of the grant and the frequency of updates.

Future application

The partners plan to use the item-level holdings analysis for future de-duplication of their serial holdings at ReCAP. This process will require further planning to determine workflows, specifically to establish the best volume for retention and reviewing best practices for deaccessioning items from the collection.

The item-level holdings analysis will also allow partners to strengthen commitments to the retention of complete print serial runs for themselves and the larger scholarly community.

In 2019, Harvard University formally joined ReCAP as a full partner and integration of Harvard's collection into the Shared Collection is ongoing. Thus, the initial analysis of the ReCAP partner's shared serials is conducted between Columbia University, NYPL, and Princeton University. The workflows and methodologies created for the item-level holdings analysis may be used if and when Harvard University integrates serials into the Shared Collection.

Conclusion

While this project is ongoing, the ReCAP partner libraries anticipate future as of now unknown use cases and applications for the item-level holdings analysis of their serials collections. This project strengthens the commitment to the shared collection model and ongoing collaborative collection development projects. In addition, it provides the partners with the opportunity to be both good stewards of the materials in their collections by better understanding the extent of their holdings, and better collection space managers as the partners explore ways in which they can be more efficient in collection storage.

This project has necessitated collaborative problem solving and coordination of decision making between the partner libraries and CRL. Working across four large institutions to prioritize and plan the analysis established protocols and workflows that may be applied to future ReCAP partner initiatives. Establishing priorities and tools for collaboration are integral to the ongoing success of this project.

Most importantly, this analysis will allow the ReCAP partners to provide detailed information about the serial titles and their completeness to which they are dedicated to preserving.

Acknowledgments

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

References

[i] Print Archive Network (PAN) Forum webpage http://www.crl.edu/programs/print-archive-network-forum-pan

[ii] Print Archives Preservation Registry (PAPR) webpage http://papr.crl.edu

[iii]Critical Corpus report summary http://www.crl.edu/sites/default/files/event_materials/Critical%20Corpus%20Analysis%20Status%20Report.pdf

[iv] OCLC Connexion client documentation https://help.oclc.org/Metadata_Services/Connexion/Connexion_client_documentation