12-13-2019

Matching Made in Heaven: Collections and Metadata Collaboration for Print Preservation

Alie Visser

Erin Johnson

Christina Zoricic

Matching Made in Heaven: Collections and Metadata Collaboration for Print Preservation

Alie Visser, (avisser9@uwo.ca) Metadata Management Librarian, Western Libraries, Western University,

Erin Johnson, (ejohns83@uwo.ca) Metadata Management Librarian, Western Libraries, Western University,

Christina Zoricic, (czorici@uwo.ca) Head, Discovery, Description, & Metadata, Western Libraries, Western University

Abstract

Following the trend of repurposing library space to meet modern user needs, Western University is undergoing a planned revitalization and renovation of its largest library on campus. As a result, 500,000 items will need to be shifted to other locations or off-site storage. In this session we will outline the impact of metadata work in shifting this large collection of material to a shared print preservation storage facility, in coordination with Western University's Keep@Downsview partnership (https://downsviewkeep.org/). Keep@Downsview is a partnership of five universities to preserve the scholarly record in Ontario in a shared, high-density storage and preservation facility.

We will demonstrate the importance of collaboration and communication between Collections Librarians and Metadata Librarians to improve identification of materials for shared print preservation. While past Charleston conference presentations have discussed weeding legacy print collections, this session will focus on the importance of metadata matching processes. Speaking from experience at Western University, we will identify the types of tools and skills that we use to facilitate this work (such as MarcEdit, Excel, Python, OpenRefine, Google Sheets, and regular expressions). In highlighting the value of metadata for collections based projects, attendees will walk away with talking points to advocate for quality metadata at their institution and with vendors.

Background

This paper summarizes Western University's shared print preservation program, Keep@Dowsnview, and the metadata work involved in shifting low use material. Located in London, Ontario, Canada, Western University (The University of Western Ontario) is an Association of Research Libraries (ARL) member with approximately 36,000 FTE. The university has seven campus libraries and three affiliated university college libraries. There are four physical storage locations for material including two on campus sites and two off campus, which from the user perspective, displays as "Storage - Use Request Item". Western University's current acquisition budget is approximately $15 million, which supports the research, teaching, and learning mission of the university.

What is Keep@Downsview?

Keep@Downsview is a shared single copy print preservation partnership between five ARL member universities in Ontario: Queen's University in Kingston, the University of Ottawa, McMaster University in Hamilton, Western University in London, and the University of Toronto. The partnership's intent is "to preserve the scholarly record in Ontario in a shared high-density storage and preservation facility located at the University of Toronto's Downsview Campus in

North Toronto. Preserving and maintaining this valuable collection ensures that these resources will be available for generations to come" (Keep@Downsview, 2017). As the partnership is not a national or provincial consortial project, each institution operates its own Integrated Library System (ILS) software. The initial proposal included the use of existing institutional inter-library loan (ILL) programs to transfer physical materials or provide desktop delivery (Horava, et. al, 2017). To eliminate duplication of content, each institution must compare and match their own bibliographic records against those held at Downsview, and subsequently also compare it to material held in the University of Toronto Library System.

Without a middleware software solution in place, the task of comparing local holdings to those of the University of Toronto has been a challenge for the partners. This initiative is currently a very labour intensive process, as metadata is compared largely via Excel spreadsheets, based on data exported from different ILS systems. As part of a consortia initiative with the Ontario Council of University Libraries (OCUL), on December 12th, 2019, three of the five partners (University of Ottawa, Queen's University, and Western University) will launch a new shared Library Services Platform (LSP) - Ex Libris' Alma. The use of a shared system will enable these three partners to view each other's holdings more efficiently, however, without the University of Toronto's holdings included in the shared LSP, metadata matching workflows for the Keep@Downsview project remain challenging.

Importance of Quality Metadata

It is critical to have quality metadata when embarking on a large scale matching project, such as Keep@Dowsnview. Yet, often this is not the case with large library collections. Material description and encoding practices have varied dramatically over the decades, leading to inconsistencies in data. As Horava, et al. emphasizes, "do not underestimate the data challenges caused by heterogeneous systems in place at different institutions" (2017). Some of the factors that may contribute to inconsistencies in cataloguing across institutions include: variation in descriptive practices, local policies and cataloguing exceptions, vendor derived records, brief records, format-blind records, unintentional typos, and "cataloguer's judgement". Metadata is messy. Each set or batch of records can contain any variety of the above listed inconsistencies. This means that each project at each library will have unique metadata challenges based on the context in which the metadata exists.

In most cases, material at Western that was selected for shared print preservation included older volumes with minimal level metadata, and no standard identifiers (such as an OCLC number or an ISBN). Additionally, the quality of metadata often varied dramatically as it included material described in a variety of formats (AACR, AACR2, RDA, etc.). Many of the records identified for the matching workflow were brief records, manually entered into an electronic system from print-based card catalog information. Taking this into account, matching to the Keep@Downsview partners becomes extremely difficult without standard match points.

OCLC Data Sync

An OCLC Data Sync (formerly, OCLC Reclamation) is a valuable step in bibliographic metadata matching since it increases the match rates in automated processes. The purpose of OCLC numbers in local records is to match local holdings with those in the world's largest Online Public Access Catalog (OPAC) to provide a common reference key for bibliographic metadata worldwide. Essentially, an OCLC Data Sync is a service that libraries can use to synchronize their local holdings to exact items in OCLC's database. OCLC assigns a new number for material it does not find an "exact match" for, provided the record meets the minimum standards

for WorldCat records. When complete, OCLC returns files with a standardized number inserted in the 035 MARC field. For records that were not matched to OCLC's database and did not meet OCLC requirements, the library receives reports identifying why each record was rejected. Two main issues within unmatched records are sparse coding and encoding errors. The Data Sync process can be complicated, confusing, and a shared struggle that many technical services units experience (see Appendix A for a selected list of related resources).

The Keep@Downsview project coincided with the previously mentioned multi-year initiative to migrate to a consortially shared Library Services Platform (LSP). There are 14 Ontario University Libraries participating in this initiative, including three from Keep@Downsview. The data migration for this new shared initiative required an OCLC Data Sync of our records, as Ex Libris uses the OCLC number as the basis for matching all member records to build what they call the consortia "Network Zone". As such, the Data Sync of our records became a critical key to both our new shared LSP and the Keep@Downsview project. It is highly recommended to complete a Data Sync before joining projects of this scale.

Match Points

The Keep@Downsview partnership emphasizes the avoidance of duplication when possible. The better the metadata of both the University of Toronto Libraries (UTL) and its partners, the greater the likelihood that material is not duplicated within the UTL system. Since partners are required to match our records to those located within one of UTL's 36 library locations, as well as at the Downsview facility, partners must sort their data into three different streams. If an item we want to send to storage does not match an existing record in the UTL system, we send both the metadata (bibliographic and holdings information) as well as the physical item. If an item we want to send is identical to one found at one of the UTL locations, we do not send our record, only our physical material and corresponding holdings metadata. If an item we want to send matches to an item located in the Downsview facility, we notify UTL to modify the holdings information of selected materials. Thus, a single, verified, match point becomes invaluable when attempting to automate a matching process.

Having a quality match point means there is less data cleanup and it makes the process easier to automate. Prior to completing an OCLC Data Sync, Western Libraries did not rely on a single match point, nor did we attempt to automate any part of the process. Rather, we used a fuzzy matching combination of ISBN, title, and imprint to manually match our records to the University of Toronto's Downsview holdings. While a manual process guarantees the highest match rate, it is an incredibly labour and time intensive process and the best approach for a small selection of materials. However, when attempting to match thousands of records, a fully human mediated matching process is not financially viable. Thus, we looked to metadata tools to facilitate the human mediation and reduce time, labour, and financial efforts.

Approaches to Metadata Matching

While there are several options to facilitate automated record matching, institutional commitment of financial resources in the process can vary widely. Investigating outsourcing options to conduct the work is highly recommended for institutions embarking on a metadata matching process/partnership. OCLC Greenglass and GoldRush are examples of products/tools that can be used to facilitate collection analysis and metadata matching. Given the "home-grown" nature of the Keep@Downsview partnership, and the time constraints of the project, Western Libraries chose to move forward with in-house Metadata Librarian driven matching. As previously noted, when exploration into the Keep@Downsview partnership began, attention was

focused on manually and visually matching records based on ISBN, title, publication information, and imprint dates. As progress was made in the collection shifting project, team member skills needed enhancement in order to improve processes. (See Appendix B for a list of tools and learning resources to help develop skills needed to effectively undertake and automate a metadata matching project.) Three approaches to metadata matching were used: visual matching (fully human mediated), Excel VLOOKUP (semi-human mediated), and a Python script (faster, semi-human mediated).

Visual Matching

{ED: Place figure 1 here. Caption: "Figure 1: Visual matching method used at Western University."}

For visual matching of ISBNs basic Excel tools such as the sort, filter, and colouration of duplicates were used. Advanced Excel add-in programs, such as ASAP Utilities and Ablebits, were also used. In visual matching, a comparison is made of two records to verify if they are exact matches by combining records from both institutions into one spreadsheet, sorting by ISBN number, and then comparing the descriptive metadata of each line. Visual matching at the bibliographic level becomes problematic for multiple reasons. First, multiple volumes of a work (such as a multi volume set, for example) can have the same ISBN number thus triggering the need for further investigation. Second, ISBNs can be used multiple times and applied (erroneously) to the print and electronic records, requiring further investigation to determine whether or not one is looking at a true match. Overall, while extremely accurate, this process is incredibly time consuming and not recommended for large data sets. However, for this project, it helped the Metadata Librarians become familiar with the metadata in order to quickly identify inconsistencies and trends.

Excel VLOOKUP

{ED: Place figure 2 here. Caption: "Figure 2: VLOOKUP matching method used at Western University."}

Once familiar with the metadata, the basic excel function - VLOOKUP - became a useful tool in semi-automating the matching process. Although semi-automated, the function still needed to be written for each batch of records. Additionally, using the VLOOKUP function meant that the data within the field needed to be perfectly clean: free of extraneous qualifiers (such as price or format) and spaces that cataloguers often add to the MARC 020 field. When a VLOOKUP is used, a manual quality check is required to remove false matches.

Python Script

{ED: Place figure 3 here. Caption: "Figure 3: Python matching method used at Western University."}

Python programming is a valuable skill for Metadata Librarians. Although it can require a significant time commitment up front, knowledge of a programming language is a skill that may be drawn upon for a wide variety of metadata projects. At Western Libraries, a simple Python script was used to automate the matching processes for OCLC and ISBN numbers by automating a VLOOKUP style function on prepared datasets. This script is linked in Appendix B under the Tools section. In order to use the script effectively, a clean match point is required as well as a data file that follows a strict set of formatting rules. These rules need to be thoroughly

documented, and the documentation needs to accompany the script. While this script is time effective in that it speeds up the matching process, a manual quality check still needs to be completed as part of the workflow.

Metadata Matching Workflow

{ED: Place figure 4 here. Caption: "Figure 4: Metadata matching workflow used at Western University."}

The metadata matching workflow used by Western Libraries follows five high level phases (See figure 4). In all cases of metadata matching, the first step is to analyze and become familiar with the data, identifying which fields would be good candidates for match points. Next, the data must be cleaned and prepared for matching (e.g. removal of extraneous spaces, data points on individual rows, etc.). Consideration must be given to end needs and output. This phase is a time consuming and unavoidable element of metadata matching, though the relative time spent on this phase can be mitigated by ensuring quality metadata from the start. The next phase of the workflow is the matching process, which is the only phase that can be automated and which Western Libraries has worked to refine. Following this phase is the important quality check, where data is reviewed by a Metadata Librarian in order to remove any false matches that may have been generated due to low quality or inaccurate data. The final phase is to sort the data into files that can be used by collections maintenance staff to physically move the material.

Communication and Collaboration

There are many layers of communication and collaboration at stake in a metadata matching project of this calibre, both at the institution and within the Keep@Downsview partnership. At both the institutional and partnership level, Metadata Librarians need to communicate with colleagues in Collections about metadata matching challenges encountered. At an institutional level, there should be clear lines of communication between Collections and Metadata Librarians. An example would be the inclusion of Metadata representation on Collections project teams and vice versa (Darcovich, Flynn & Li, 2019). It is important to acknowledge the interconnectedness between the work of Metadata and Collections Librarians and to act accordingly, so material is handled appropriately throughout the collection lifecycle. Developing these stronger communication channels between departments will help projects run smoothly from start to finish (van Ballegooie, 2015).

It is crucial that staff involved in a project like Keep@Downsview have a basic understanding of metadata and collections processes, and the significance of both throughout the project life cycle. Inviting staff to an initial project orientation meeting where they are provided with an overview of the project can set the tone for success. Previous studies, such as those by Darcovich, Flynn, and Li (2019), have noted that when staff have limited metadata and cataloguing knowledge, this can later affect metadata clean up, and in the case of Western Libraries' participation in Keep@Downsview: metadata matching.

Another area in which collaborative communication may improve project outcomes centers on the concept of a well-defined workflow for partner institutions to follow. Creation of this workflow could include clear guidelines and procedures for metadata, best practices, and the responsibilities therein for each member of the partnership. For example, there is some discussion among librarians (Maiorana, Bogus, Miller, Nadal, Risseeuw & Hain Teper, 2019), regarding a best practice to record information on the material's condition in the MARC 583 note field to improve consortia collaboration into the future. Following this practice at large could

assist metadata librarians in matching to the best possible copy for preservation. Creating and communicating a shared workflow for collaborative projects enables the consistent use of field data, such as the MARC 583 note, when working on shared collections projects.

Next Steps: Advocating for Quality Metadata

Advocating for quality metadata locally includes consistent investment in training and professional development for staff in technical services departments. There is a strategic advantage to these investments, as high quality metadata creation and management relies on people with specialized knowledge, skills, and experience. Tasks like large scale print preservation projects that are based on metadata matching can encounter significant metadata barriers, which can reduce the overall quality and rate of matching and jeopardize projects. These barriers may be lessened for future initiatives by adequately maintaining metadata standards and description, a model that requires skilled technical services staff. A significant investment in people is an investment in quality metadata, which has the potential to mitigate data-related roadblocks and save on future labour costs.

Quality metadata also requires building strong relationships and communication channels with vendors. It is important to periodically evaluate the quality of vendor supplied bibliographic data to ensure it meets minimum standards, as outlined in a vendor agreement (if applicable). Different agreements offer different levels of cataloguing, so it is important for Metadata Librarians to be consulted on these elements of a contract. In advocating for quality metadata, Collections Librarians should form a feedback loop between the metadata team and vendors to ensure improvements in record quality are communicated and realized. Communication and teamwork at the onset can go a long way to aiding in the success of long term collaborative metadata and shared print preservation projects.

Appendix A - Selected OCLC Data Sync Resources

OCLC Data Sync Collections support documentation - https://help.oclc.org/Metadata_Services/WorldShare_Collection_Manager/Choose_your_Collection_Manager_workflow/Data_sync_collections

ALCTS Presentation: Surviving An OCLC Data Sync by Kim Edwards, George Mason University, June 2018 https://drive.google.com/file/d/1B6tNnH-YVKGdhIH7PQLxKljd3IbwIp-t/view

GALILEO Interconnected Libraries Documentation - OCLC Data Sync https://sites.google.com/view/g3almatraining/special-projects/oclc-data-sync

Harvard University - OCLC Data Sync (Maintaining Holdings in WorldCat) - https://wiki.harvard.edu/confluence/pages/viewpage.action?pageId=229314649

Appendix B - Metadata Toolkit and Learning Resources

Tools

MarcEdit - https://marcedit.reeset.net/

Ablebits - https://www.ablebits.com/
ASAP Utilities - https://asap-utilities.com/
OpenRefine - http://openrefine.org/
Regular Expressions - https://www.regular-expressions.info/
Python - https://www.python.org/
Keep@Downsview Metadata Matching Script -
https://github.com/ernieejo/downsviewmetadatamatching

Learning Resources

Terry Reese YouTube Channel - https://www.youtube.com/user/tpreese
MarcEdit Development Website - https://marcedit.reeset.net/
Library Carpentry: Open Refine - https://librarycarpentry.org/lc-open-refine/
Library Carpentry: Python Intro for Libraries - https://librarycarpentry.org/lc-python-intro/aio.html
Automate the Boring Stuff with Python - https://automatetheboringstuff.com/
Lynda.com - https://www.lynda.com/
Improve your Excel - http://www.improveyourexcel.com/
Reddit /rLearnPython - https://www.reddit.com/r/learnpython/
Stackoverflow - https://stackoverflow.com/

References

Darcovich, J., Flynn, K., & Li, M. (2019). Born of collaboration: the evolution of metadata standards in an aggregated environment. VRA Bulletin, 45(2), 1–12. Retrieved from https://online.vraweb.org/vrab/vol45/iss2/5

Horova, Tony; Rykse, Harriet; Smithers, Anne; Tillman, Caitlin; and Wyckoff, Wade. Making Shared Print Management Happen: A Project of Five Canadian Academic Libraries. (2017). Western Libraries Publications. 58. https://ir.lib.uwo.ca/wlpub/58

Keep@Downsview. (2017). Retrieved November 22, 2019, from https://downsviewkeep.org/home

Maiorana, Z., Bogus, I., Miller, M., Nadal, J., Risseeuw, K., & Teper, J. (2019). Everything Not Saved Will Be Lost: Preservation in the Age of Shared Print and Withdrawal Projects. College & Research Libraries, 80(7), 945. doi:https://doi.org/10.5860/crl.80.7.945

Panchyshyn, R. S. (2012). Benefits of Batch Reclamation: The Kent State University Libraries Experience. Cataloging & Classification Quarterly, 50(1), 3–16. https://doi.org/10.1080/01639374.2011.622836.

Thornburg, Gail, and W. Michael Oskins. (2007). Misinformation and bias in metadata processing: matching in large databases. Information Technology and Libraries, 26(2), 15-25. https://doi.org/10.6017/ital.v26i2.3278.

van Ballegooie, M., & Borie, J. (2015). Facing Our E-Demons: The Challenges of E-Serial Management in a Large Academic Library. The Serials Librarian, 68, 342–352. Retrieved from https://www.tandfonline.com/doi/full/10.1080/0361526X.2015.1017714.