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

Erin Johnson
Alie Visser
Christina Zoricic

Follow this and additional works at: https://ir.lib.uwo.ca/wlpres
Part of the Cataloging and Metadata Commons, and the Collection Development and Management Commons
Matching Made in Heaven

Collections and metadata collaboration for print preservation
Who?

Erin Johnson
ejohns83@uwo.ca
Twitter: @erinee_jo

Alie Visser
avisser9@uwo.ca

Christina Zoricic
czorici@uwo.ca
Twitter: @Libraried
You’re from where?

Western University
London, Ontario, CANADA
www.westernu.ca
Western University

~36,000 FTE

~$15.4 million acquisitions budget

7 campus libraries; 3 affiliated university college libraries

4 physical storage locations all appear as “Storage” to the user

- External, contracted offsite
- RDL (essentially dark storage)
- Archives and Research Collections Centre (ARCC)
- Keep@Downsview (new)
Agenda

Introduction to the Keep @ Downsview project

Metadata quality and its importance

Match points and the tools/skills used in metadata matching

Why it’s important to communicate and collaborate with metadata team

Talking points to advocate for good metadata records
Shared single copy print preservation partnership between 5 ARL member Ontario Universities.

Currently 5 universities all on separate ILS systems trying to match bibliographic data to one location.

https://downsviewkeep.org/
Metadata quality and its importance

- Projects such as this require quality metadata:
  - Garbage in = Garbage out

- OCLC Data Sync (aka Reclamation)
  - Synchronizes local holdings with those in WorldCat.
  - OCLC numbers are input into all bibliographic records.

- Why OCLC numbers are important:
  - It’s a standard identifier used by libraries worldwide.
  - Commonly used by consortia to match records in shared discovery environments.
  - Aids in matching processes!
Match Points in Metadata

Unique reference keys commonly used between the records being matched

Eg. ISBN/ISSN, OCLC number

Good Match Points

= Less Data Clean Up
Easier to Automate
Matching Challenges: ISBN

- Extraneous data in the 020 field = messy match point
- Voluming (ISBN for set vs. ISBN for different volumes in a set)
- Not included in all records

<table>
<thead>
<tr>
<th>020</th>
<th>__</th>
<th>a 9780198569961</th>
<th>q (pbk.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>020</td>
<td>__</td>
<td>a 9780198569954</td>
<td>q (hbk.)</td>
</tr>
<tr>
<td>020</td>
<td>__</td>
<td>a 0198569963</td>
<td>q (pbk.)</td>
</tr>
<tr>
<td>020</td>
<td>__</td>
<td>a 0198569955</td>
<td>q (hbk.)</td>
</tr>
</tbody>
</table>
Matching Challenges: OCLC Control Number

- Currency of OCLC synchronization
- Legacy data decisions
- Not included in all records

<table>
<thead>
<tr>
<th>035</th>
<th>$$a (CA-ONBEC)946315-01ocul_nip</th>
</tr>
</thead>
<tbody>
<tr>
<td>035</td>
<td>$$a (SIRSI)u946315 $$9 ExL</td>
</tr>
<tr>
<td>035</td>
<td>$$a (Sirsi)#o843198462</td>
</tr>
<tr>
<td>035</td>
<td>$$a (OCoLC)843198462</td>
</tr>
</tbody>
</table>
Library records lack uniformity

- Brief vs. full records
- Local variation
- Unintentional variation i.e. typographical errors
- Format-blind records
- Created under different schema
‘Do not underestimate the data challenges caused by heterogeneous systems in place at different institutions…. Different cataloging practices impact how items can be searched, matched, and disposed.’

(Horava, et al., 2017)
ALL DATA IS MESSY!

- ‘Best’ match point is contextual to the datasets being matched
- Always data clean up to prepare for the matching process
Approaches to Metadata Matching

- Outsource eg. OCLC greenglass
- Visual matching
- VLOOKUP function in Excel
- Python script
Visual Matching

- Visual scan
- Manual process
- Time consuming
- Human error
Excel VLOOKUP

- Quick, accessible tool
- Semi-automated
- Requires clean match point
- Manual quality check
Python Script

- Basic programming
- Documentation needed
- Requires clean match point
- Specifics for data preparation
- Manual quality check

```python
import pandas as pd
from pandas import DataFrame, read_excel, merge, ExcelWriter

# matches against OCLC numbers
df_1 = read_excel('U:\...WL-UTL data.xlsx', sheet_name='WL-OCLC')
df_2 = read_excel('U:\...WL-UTL data.xlsx', sheet_name='UTL-OCLC')
df_3 = df_1.merge(df_2, on='OCLC', how='inner')
df_4 = read_excel('U:\...WL-UTL data.xlsx', sheet_name='WL-ISBN')
df_5 = read_excel('U:\...WL-UTL data.xlsx', sheet_name='UTL-ISBN')
df_6 = df_4.merge(df_5, on='ISBN', how='inner')
df_7 = df_3.merge(df_6, how='outer')
df_8 = df_4.merge(df_7, how='left')

# writes a new spreadsheet
writer = pd.ExcelWriter('U:\...WL-UTL match.xlsx', engine='xlsxwriter')
df_3.to_excel(writer, sheet_name='OCLC')
df_6.to_excel(writer, sheet_name='ISBN')
df_7.to_excel(writer, sheet_name='COMBINED')
df_8.to_excel(writer, sheet_name='UNMATCHED')
writer.save()
```
Metadata Management Toolkit
Spreadsheet Software

Function
- Organize and display data
- Manipulate and analyze data
- Varying functionality
  i.e. Add-ons, RegEx

Learning Resources
- Lynda.com
- Improveyourexcel.com
MarcEdit

Function
- Export and work with the MARC records of another institution
- Create and manipulate MARC records
- RegEx functionality

Learning Resources
- YouTube - tpreese channel
- MarcEdit Development Website
OpenRefine

Function
● Desktop application for data cleanup
● Parse and analyze data
● Formulas to transform data
  i.e. RegEx, GREL, Jython

Learning Resources
● OpenRefine Wiki on Github
● Library Carpentry: OpenRefine
Python Programming

Function
- Readable programming language
- Data manipulation and analysis
- Automate processes
- Software libraries to hold data sets
  i.e. Pandas

Learning Resources
- Automate the Boring Stuff
- Library Carpentry: Python Intro for Libraries
- /rLearnPython, Stackoverflow
Communication & Collaboration

How to improve communication:

- Decrease ‘silos’.
- Become advocates for quality metadata.
- Create clear lines of communication.

How to improve collaboration:

- Train staff to have a basic understanding of metadata and collections work.
- Shared inter-departmental workflows.
- Shared standards and guidelines.
Advocate for Quality Metadata: At your Institution

- Invest in technical services!
  - Strategically plan for the future.
  - People create metadata, so invest in them.
  - Trial software licensing tools.

- Reduce barriers in the future by maintaining metadata now:
  - Save $ on labour later when clean is more intensive.
  - Become an advocate.
Advocate for Quality Metadata: With your Vendors

- Ask to review analytics:
  - Do they meet minimum standards?

- Forge good relationships with your vendors
  - Periodically evaluate the quality of supplied records.
  - Different agreements = differing levels of cataloguing services.

- Communicate with your vendors:
  - Offer feedback on their supplied metadata.
  - They can’t improve if you don’t communicate.
Questions/Discussion

Thank you!

Erin Johnson
ejohns83@uwo.ca
Twitter: @erinee_jo

Alie Visser
avisser9@uwo.ca

Christina Zoricic
czorici@uwo.ca
Twitter: @Libraried
Metadata Toolkit

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@Downsv view Metadata Matching Script - https://github.com/ernieejo/downsv viewmetadatamatching
Metadata Toolkit: 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/


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/


Image Attributions


https://commons.wikimedia.org/wiki/File:Microsoft_Excel_2013_logo.svg

https://marcedit.reeset.net

https://commons.wikimedia.org/wiki/File:OpenOffice.svg

https://commons.wikimedia.org/wiki/Category:OpenRefine#/media/File:OpenRefine_New_Logo.png

https://commons.wikimedia.org/wiki/File:Python_logo_and_wordmark.svg

https://pixabay.com/vectors/protesting-megaphone-hand-woman-3411685/

https://pixabay.com/vectors/megaphone-communicate-announce-3396672/

https://www.flickr.com/photos/theilr/4806856995