11-7-2019

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

Erin Johnson

Alie Visser

Christina Zoricic

Matching Made in Heaven

Collections and metadata collaboration
for print preservation

Who?

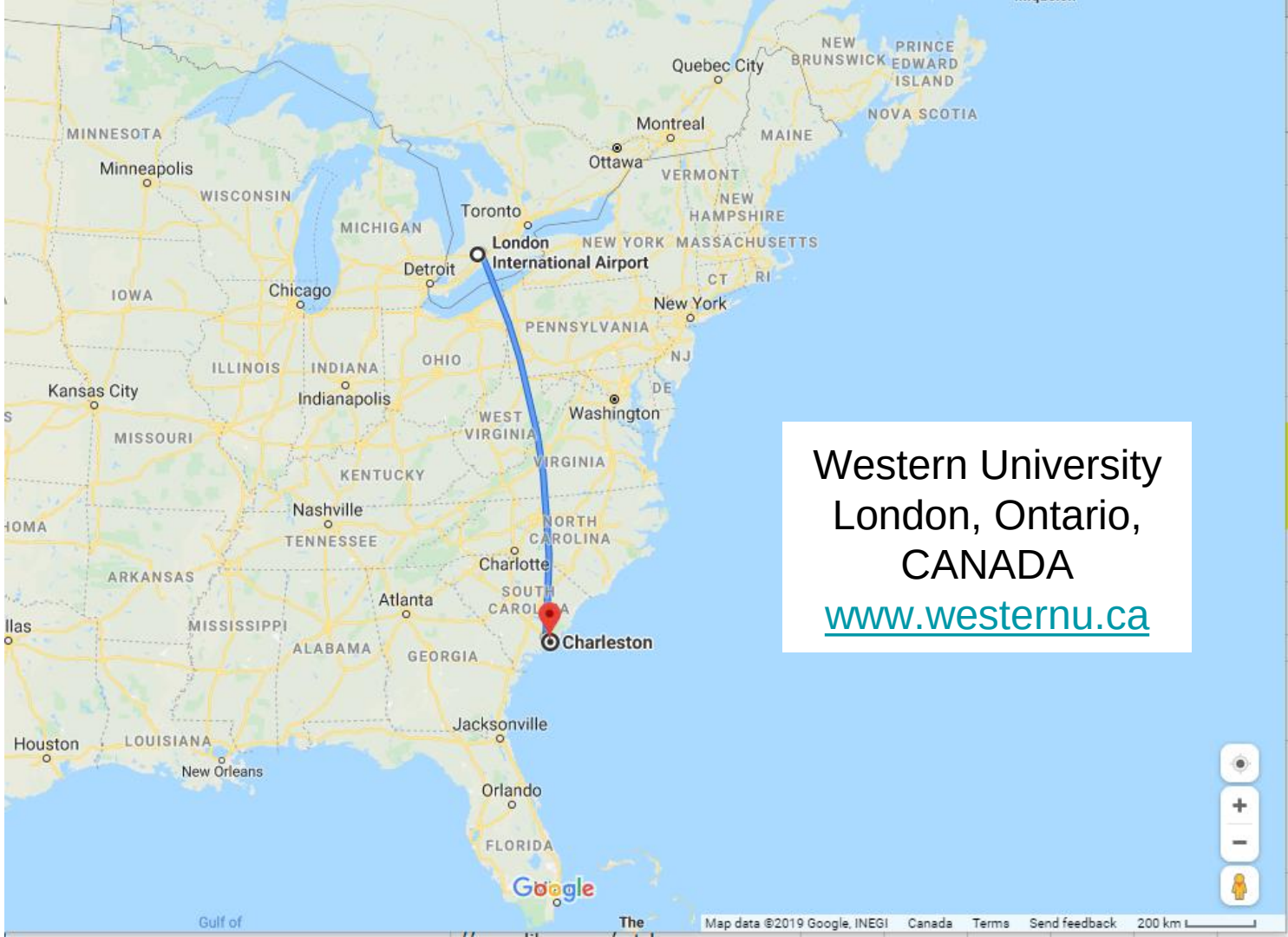
Erin Johnson
ejohns83@uwo.ca
Twitter: @erinee_jo

Alie Visser
avisser9@uwo.ca

Christina Zoricic
czorici@uwo.ca
Twitter: @Libraried

Western
UNIVERSITY · CANADA

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

Keep@Downsview

https://downsviewkeep.org/

Français
Home
About
Contact
News

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.

Western University · Canada

McMaster University

University of Toronto

Queen's University

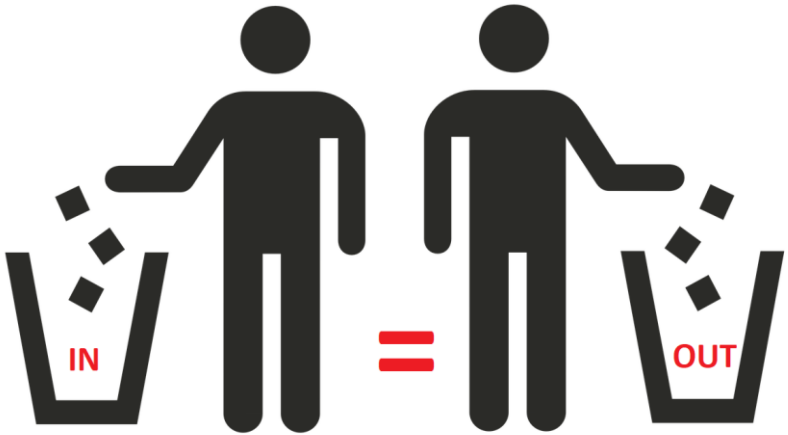
uOttawa

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

	020	__ |a 9780198569961 |q (pbk.)
	020	__ |a 9780198569954 |q (hbk.)
	020	__ |a 0198569963 |q (pbk.)
	020	__ |a 0198569955 |q (hbk.)

Matching Challenges: OCLC Control Number

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

035	$$a (CA-ONBEC)946315-01ocul_nip
035	$$a (SIRSI)u946315 $$9 ExL
035	$$a (Sirsi)#o843198462
035	$$a (OCoLC)843198462

Library records lack uniformity

- Brief vs. full records
- Local variation
- Unintentional variation ie. typographical errors
- Format-blind records
- Created under different schema

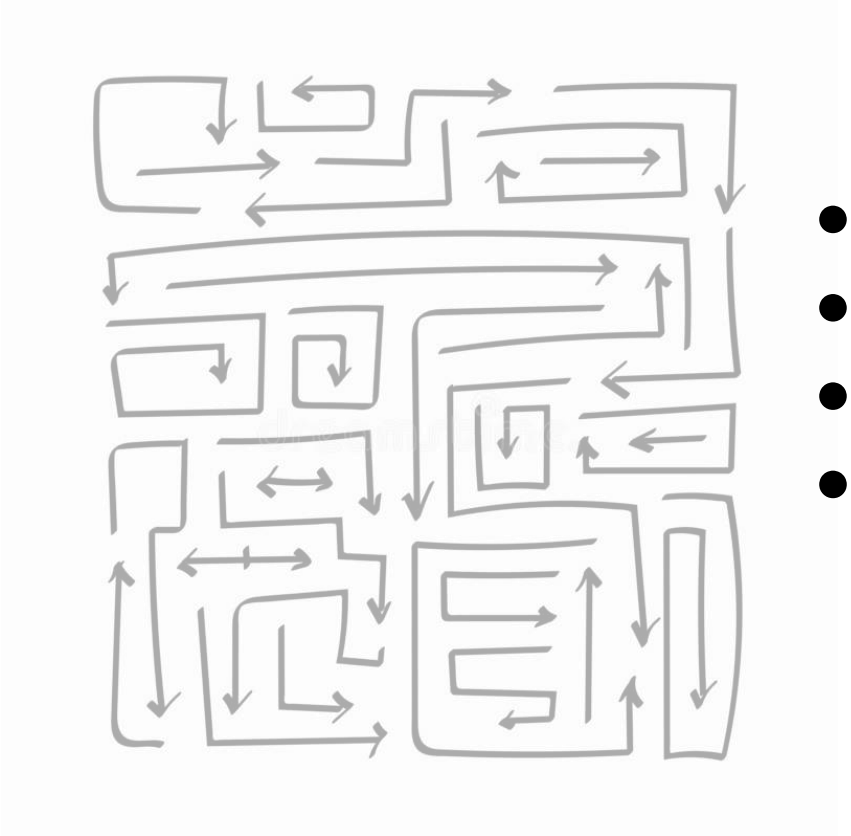
‘Do not underestimate the data challenges caused by heterogeneous systems in place at different institutions…. Different cataloging practices impact how items can be searched, matched, and disposed.’

(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**

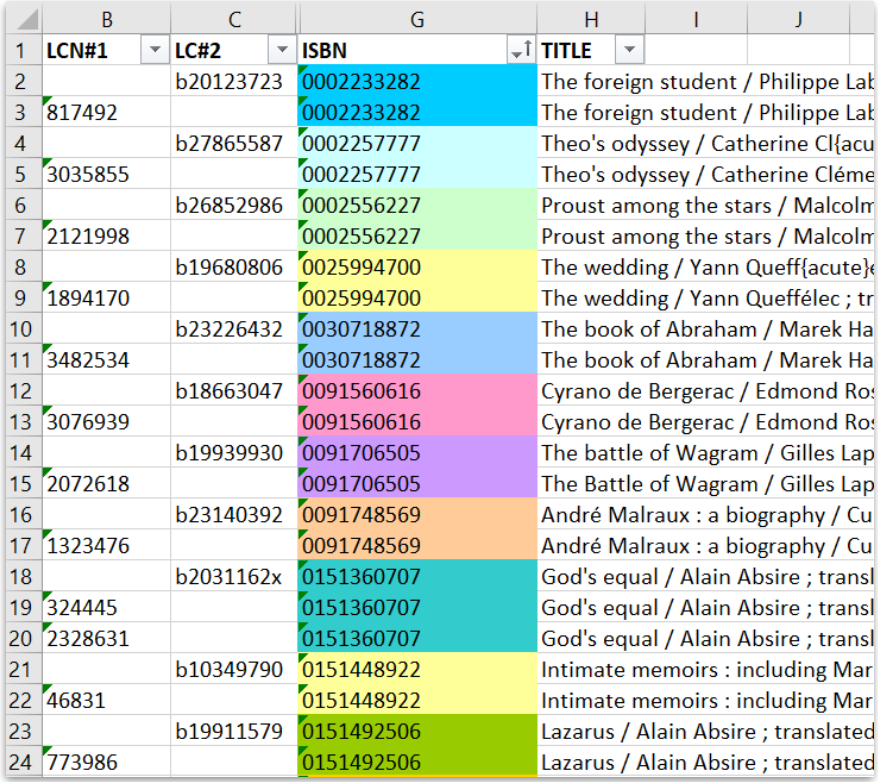
Approaches to Metadata Matching



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

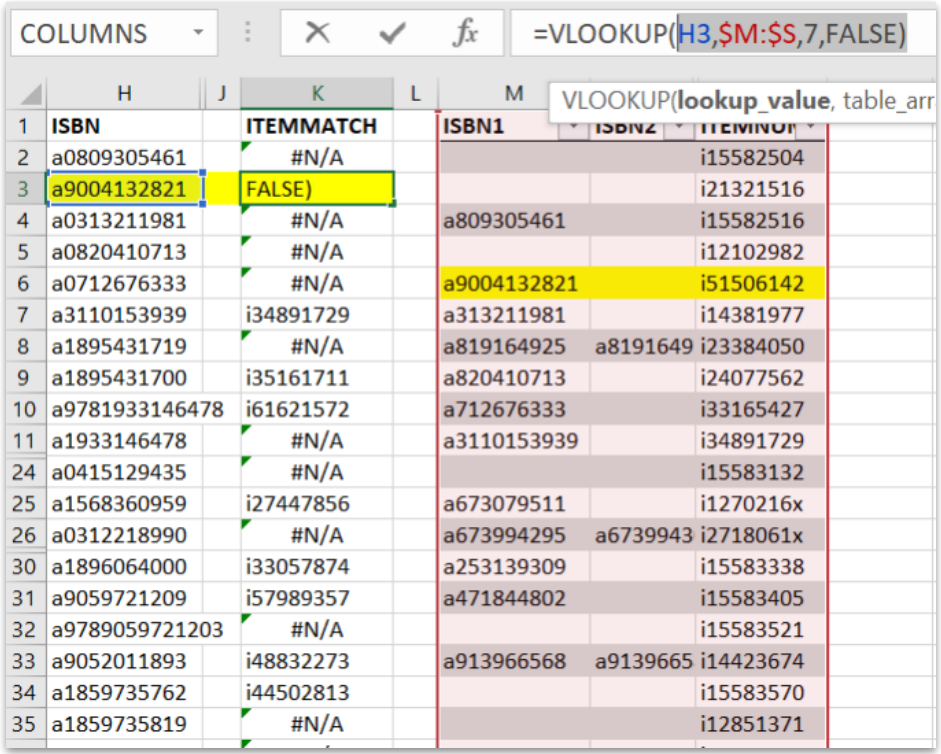
Visual Matching

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



	B	C	G	H	I	J
1	LCN#1	LC#2	ISBN	TITLE		
2		b20123723	0002233282	The foreign student / Philippe Lab		
3	817492		0002233282	The foreign student / Philippe Lab		
4		b27865587	0002257777	Theo's odyssey / Catherine Cl{acu		
5	3035855		0002257777	Theo's odyssey / Catherine Cléme		
6		b26852986	0002556227	Proust among the stars / Malcolm		
7	2121998		0002556227	Proust among the stars / Malcolm		
8		b19680806	0025994700	The wedding / Yann Queff{acute}é		
9	1894170		0025994700	The wedding / Yann Queffélec ; tr		
10		b23226432	0030718872	The book of Abraham / Marek Ha		
11	3482534		0030718872	The book of Abraham / Marek Ha		
12		b18663047	0091560616	Cyrano de Bergerac / Edmond Ros		
13	3076939		0091560616	Cyrano de Bergerac / Edmond Ros		
14		b19939930	0091706505	The battle of Wagram / Gilles Lap		
15	2072618		0091706505	The Battle of Wagram / Gilles Lap		
16		b23140392	0091748569	André Malraux : a biography / Cu		
17	1323476		0091748569	André Malraux : a biography / Cu		
18		b2031162x	0151360707	God's equal / Alain Absire ; transl		
19	324445		0151360707	God's equal / Alain Absire ; transl		
20	2328631		0151360707	God's equal / Alain Absire ; transl		
21		b10349790	0151448922	Intimate memoirs : including Mar		
22	46831		0151448922	Intimate memoirs : including Mar		
23		b19911579	0151492506	Lazarus / Alain Absire ; translated		
24	773986		0151492506	Lazarus / Alain Absire ; translated		

Excel VLOOKUP



	H	J	K	L	M			
1	ISBN		ITEMMATCH		ISBN1	ISBN2	ITEMNUM	
2	a0809305461		#N/A				i15582504	
3	a9004132821		FALSE)				i21321516	
4	a0313211981		#N/A		a809305461		i15582516	
5	a0820410713		#N/A				i12102982	
6	a0712676333		#N/A		a9004132821		i51506142	
7	a3110153939		i34891729		a313211981		i14381977	
8	a1895431719		#N/A		a819164925	a8191649	i23384050	
9	a1895431700		i35161711		a820410713		i24077562	
10	a9781933146478		i61621572		a712676333		i33165427	
11	a1933146478		#N/A		a3110153939		i34891729	
24	a0415129435		#N/A				i15583132	
25	a1568360959		i27447856		a673079511		i1270216x	
26	a0312218990		#N/A		a673994295	a6739943	i2718061x	
30	a1896064000		i33057874		a253139309		i15583338	
31	a9059721209		i57989357		a471844802		i15583405	
32	a9789059721203		#N/A				i15583521	
33	a9052011893		i48832273		a913966568	a9139665	i14423674	
34	a1859735762		i44502813				i15583570	
35	a1859735819		#N/A				i12851371	

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

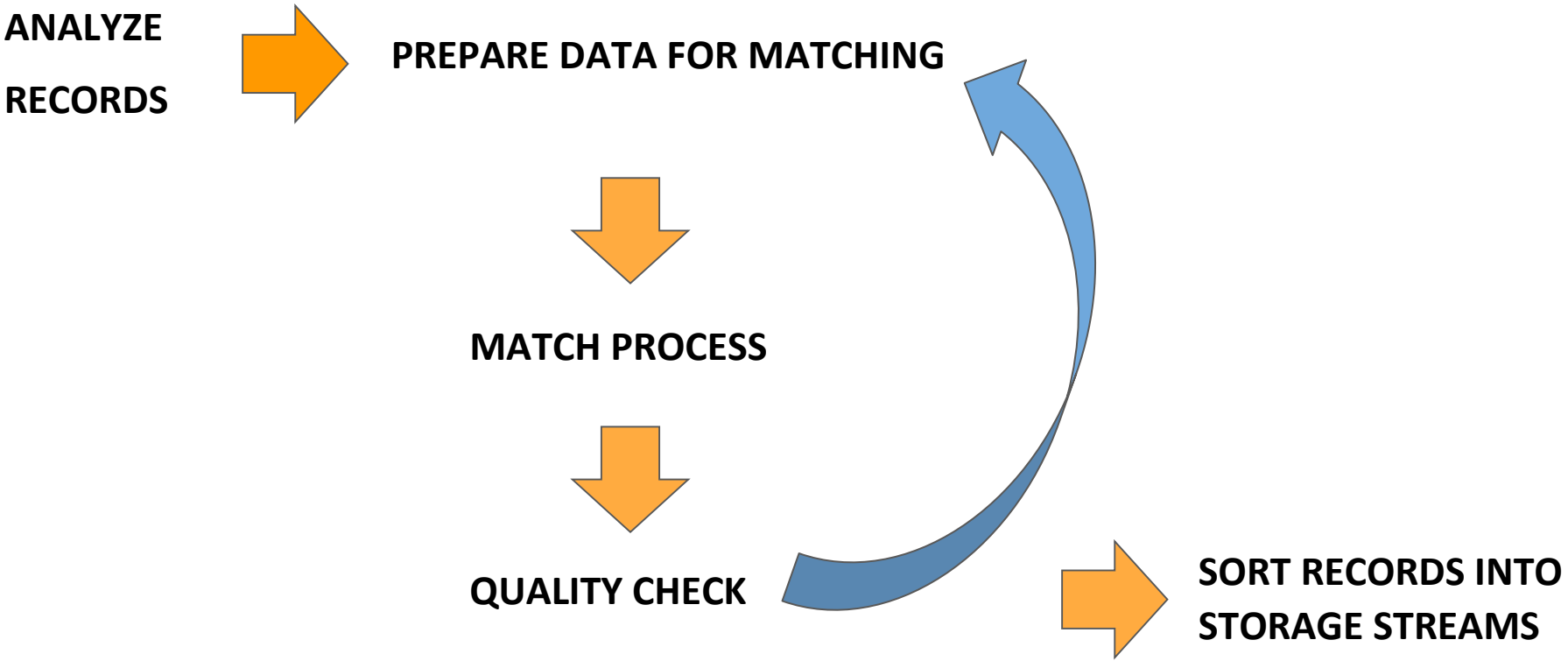
Python Script

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

21 lines (18 sloc) | 858 Bytes

```
1  import pandas as pd
2  from pandas import DataFrame, read_excel, merge, ExcelWriter
3
4  #matches against OCLC numbers
5  df_1 = read_excel('U:\\...WL-UTL data.xlsx', sheet_name='WL-OCLC')
6  df_2 = read_excel('U:\\...WL-UTL data.xlsx', sheet_name='UTL-OCLC')
7  df_3 = df_1.merge(df_2, on='OCLC', how='inner')
8  df_4 = read_excel('U:\\...WL-UTL data.xlsx', sheet_name='WL-ISBN')
9  df_5 = read_excel('U:\\...WL-UTL data.xlsx', sheet_name='UTL-ISBN')
10 df_6 = df_4.merge(df_5, on='ISBN', how='inner')
11 df_7 = df_3.merge(df_6, how='outer')
12 df_8 =df_4.merge(df_7, how='left')
13
14 # writes a new spreadsheet
15 writer = pd.ExcelWriter('U:\\...WL-UTL match.xlsx', engine='xlsxwriter')
16 df_3.to_excel(writer, sheet_name='OCLC')
17 df_6.to_excel(writer, sheet_name='ISBN')
18 df_7.to_excel(writer, sheet_name='COMBINED')
19 df_8.to_excel(writer, sheet_name='UNMATCHED')
20 writer.save()
```

Metadata Matching Workflow



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@Downsview Metadata Matching Script - https://github.com/ernieejo/downsviewmetadatamatching

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/

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

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.

Image Attributions

Anonymous. (n.d.). Trash, environment, garbage, rubbish image. Retrieved from https://pixabay.com/images/id-310219/

https://www.publicdomainpictures.net/en/view-image.php?image=260633&picture=mixed-gears

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/File:Google-Sheets-Logo-2019.png

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