Yale

OpenRefine

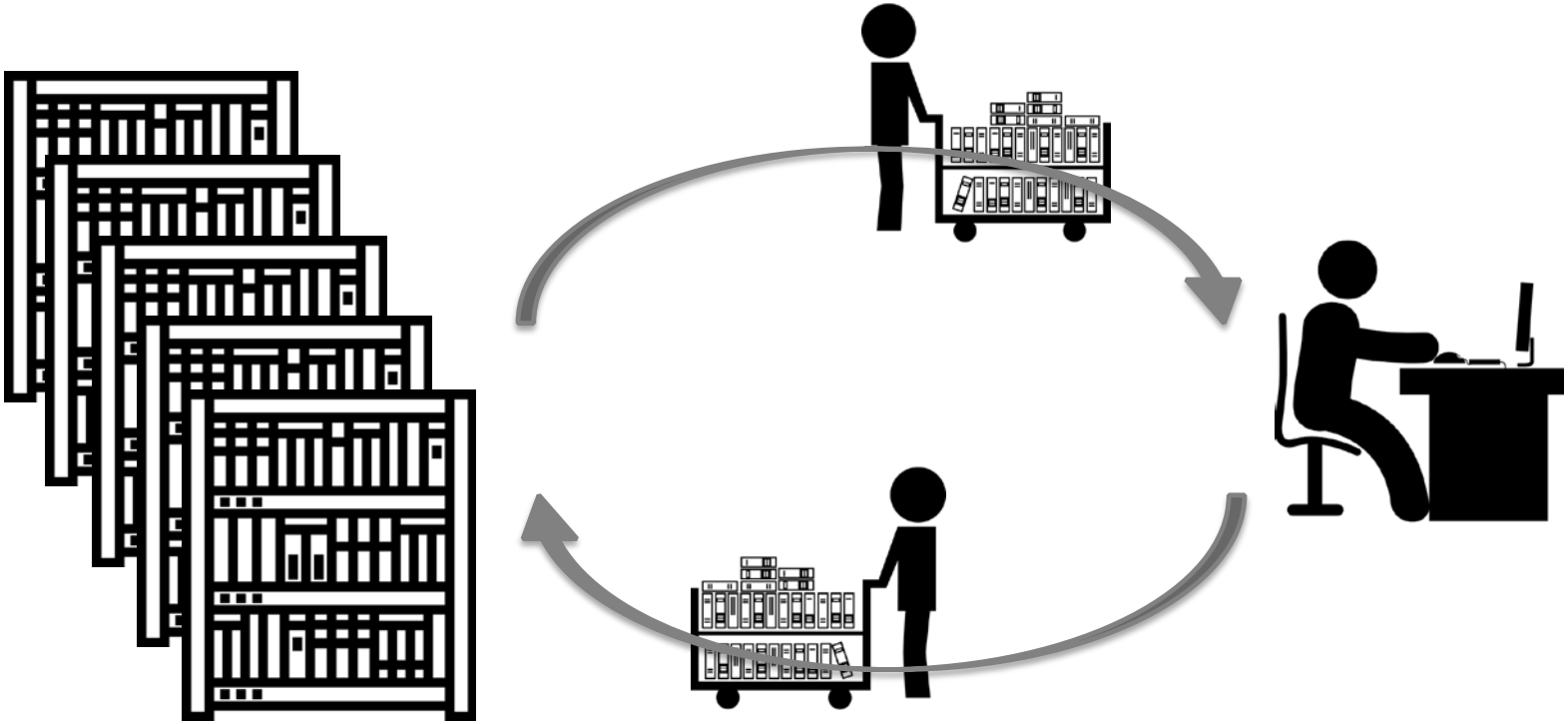
for automating backlog records searching

June 25, 2018

Yukari Sugiyama and Debbie Falvey

Yale University Library

Why automate backlog searching?



What is OpenRefine?

- Tool for cleaning and enriching data
- Free
- Browser-based web application
    - Works best in Chrome and Firefox
    - Doesn’t work in IE
- “Excel on steroids”
    - Ability to fetch URL
    - Ability to parse data
    - Ability to program operation

Demo

BIB_ID	MFHD_ID	TITLE_BRIEF	BEGIN_PU	LANGUAG	DISPLAY_CALL_NO	LCCN	ISBN	NETWORK_NUMBER
4038823	4405235	Greatest Superman stories ever told.	1995	eng	UNCAT008176		1852865954 (pbk) :	(OCoLC)ocm33278093
4633817	5025559	Ecriture de la ruse /	2000	fre	UNCAT127293	00418550	9042015713	(OCoLC)ocm44770187
4635677	5027449	Legend of Daniel Williams,	1958	eng	DIVUNCAT4635677			(OCoLC)ocm03755470
4688289	5081711	Power game : Fianna Fail since Lemass /	2000	eng	UNCAT141182	00455443	086278588X	(OCoLC)ocm44603235
4732988	5127559	Xin Zhongguo mao yi si xiang shi /	1999	chi	UNCAT182186	2001420544	7810493590	(OCoLC)ocn793054015
6263085	6928559	Tshisekedi et le destin du Congo /	2003	fre	UNCAT271409			(OCoLC)71071850
6323178	6984759	Mediance paysagere /	2002	fre	UNCAT792240			
6350307	7007651	Lao Xianggang : dong fang zhi zhu /	2000	chi	CHIUNCAT297393	2002325146	7534411432	(CStRLIN)DCLP02-B328
6350323	7007661	Lao Hangzhou : hu shan ren jian /	2000	chi	CHIUNCAT297395	2002325148	7534411483	(CStRLIN)DCLP02-B331
6451129	7097003	Pretres subversifs /	2002	fre	UNCAT903963	2004430714	2884643702 (pbk.)	(DLC) 2004430714
6478572	7118953	Saint Pierre Claver, apotre des negres.	1893	fre	DIVUNCAT6478572			(OCoLC)ocm00830370
6526807	7159007	Carte administrative de l'Algerie.	2003	fre	NEAUNCAT303272	2004335002		(OCoLC)ocm55125105
6582350	7204587	"Odes a ma douche" /	1986	fre	UNCAT799380			
6622319	7236347	Myanmar laws, 1993-1994.	2003	eng	UNCAT912586	2003333327		(CStRLIN)DCLC2003333327-B
6718623	7317035	Hong Fangzhou yan jiu lun ji /	1998	chi	CHIUNCAT360891			(OCoLC)43682117
6721631	7319517	Sathani Kaset Luang Ang Khang.	2000	tha	THAUNCAT947598			(OCoLC)ocn223071754
6785840	7366110	Cinema spiritualiste /	2004	fre	UNCAT419393	2005397150	220407490X	(FrPJT)JTL00148655
7216911	7741214	Knife, fork, spoon /	1973	eng	UNCAT827293			
7418136	7937563	Midiya aura sahitya.		eng	SASUNCAT805552			
7422557	7941297	Parochieregisters van O.L.V.-Waver /	2000	dut	UNCAT801584			(OCoLC)ocn868153117
7857096	8309765	Chutplian prathet Thai /	2006	tha	THAUNCAT947262			(OCoLC)ocn181141209
7909097	8353529	Zeng xiang quan tu San guo yan yi /	1931	chi	CHIUNCAT557295			(OCoLC)ocm21655978
7953794	8387999	Botkhwam lang rang chut thi song /	1971	tha	THAUNCAT94670			(OCoLC)ocm63991091
7990184	8417792	Bprian phasa lanna /	19uu	tha	THAUNCAT946672			
8090080	8504859	Tawq al-yamam : riwayah /	2008	ara	NEAUNCAT562892	2008334756	9953682879	(OCoLC)ocn793073049
8153862	8559761	Lok thang bai hai nai khondieo	2001	tha	THAUNCAT946725			(OCoLC)ocn417243015
8200437	8600087	Italo Calvino : licoes de modernidade /	2007	por	UNCAT791223		9789728881429 (pbk.)	
8627322	9047820	Best of the Mighty King Kong	2007	eng	UNCAT677615	2008345655		(OCoLC)ocn276394008
8710351	9128479	Census of India 2001	uuuu	eng	SSLUNCAT358862			
12898566	12990014	Vie de bonnes fortunes /	2014	fre	UNCAT966593	2015426320	9782916868318	(OCoLC)ocn915324203


ALA_DEMO - OpenRefine
127.0.0.1:3333/project?project=1816113337615
https://youtu.be/4zoIGS1G_QI
Bookmarks
Thesaurus
GitHub
RDA Toolkit
Cataloging @ Yale
Other bookmarks
ALA_DEMO
Permalink
Open...
Export
Help
Facet / Filter
Undo / Redo 0
Using facets and filters
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.
Not sure how to get started?
Watch these screencasts
100 rows
Extensions:
Show as: rows records
Show: 5 10 25 50 rows
« first ‹ previous 1 - 50 next › last »
All
BIB_ID
MFHD_ID
TITLE_BRIEF
BEGIN_PUB_DA
LANGUAGE
DISPLAY_CALL_
LCCN
ISBN
1. 4038823 4405235 Greatest Superman stories ever told. 1995 eng UNCAT008176 1852865954 (pbk) :
2. 4633817 5025559 Ecriture de la ruse / 2000 fre UNCAT127293 00418550 9042015713
3. 4635677 5027449 Legend of Daniel Williams, 1958 eng DIVUNCAT4635677
4. 4688289 5081711 Power game : Fianna Fail since Lemass / 2000 eng UNCAT141182 00455443 086278588X
5. 4732988 5127559 Xin Zhongguo mao yi si xiang shi / 1999 chi UNCAT182186 2001420544 7810493590
6. 6263085 6928559 Tshisekedi et le destin du Congo / 2003 fre UNCAT271409
7. 6323178 6984759 Mediance paysagere / 2002 fre UNCAT792240
8. 6350307 7007651 Lao Xianggang : dong fang zhi zhu / 2000 chi CHIUNCAT297393 2002325146 7534411432
9. 6350323 7007661 Lao Hangzhou : hu shan ren jian / 2000 chi CHIUNCAT297395 2002325148 7534411483
10. 6451129 7097003 Pretres subversifs / 2002 fre UNCAT903963 2004430714 2884643702 (pbk.)
11. 6478572 7118953 Saint Pierre Claver, apotre des negres. 1893 fre DIVUNCAT6478572
12. 6526807 7159007 Carte administrative de l'Algerie. 2003 fre NEAUNCAT303272 2004335002
13. 6582350 7204587 "Odes a ma douche" / 1986 fre UNCAT799380
14. 6622319 7236347 Myanmar laws, 1993-1994. 2003 eng UNCAT912586 2003333327
15. 6718623 7317035 Hong Fangzhou yan jiu lun ji / 1998 chi CHIUNCAT360891
16. 6721631 7319517 Sathani Kaset Luang Ang Khang. 2000 tha THAUNCAT947598
17. 6785840 7366110 Cinema spiritualiste / 2004 fre UNCAT419393 2005397150 220407490X
18. 7216911 7741214 Knife, fork, spoon / 1973 eng UNCAT827293
19. 7418136 7937563 Midiya aura sahitya. eng SASUNCAT805552
20. 7422557 7941297 Parochieregisters van O.L.V.-Waver / 2000 dut UNCAT801584
21. 7857096 8309765 Chutplian prathet Thai / 2006 tha THAUNCAT947262
22. 7909097 8353529 Zeng xiang quan tu San guo yan yi / 1931 chi CHIUNCAT557295
23. 7953794 8387999 Botkhwam lang rang chut thi song / 1971 tha THAUNCAT94670
24. 7990184 8417792 Bprian phasa lanna / 19uu tha THAUNCAT946672
25. 8090080 8504859 Tawq al-yamam : riwayah / 2008 ara NEAUNCAT562892 2008334756 9953682879
openrefine
ALA_DEMO - OpenR...
C:\Users\ys394\Desk...
Record video
EN
9:59 AM

Steps 1 – Cleaning data

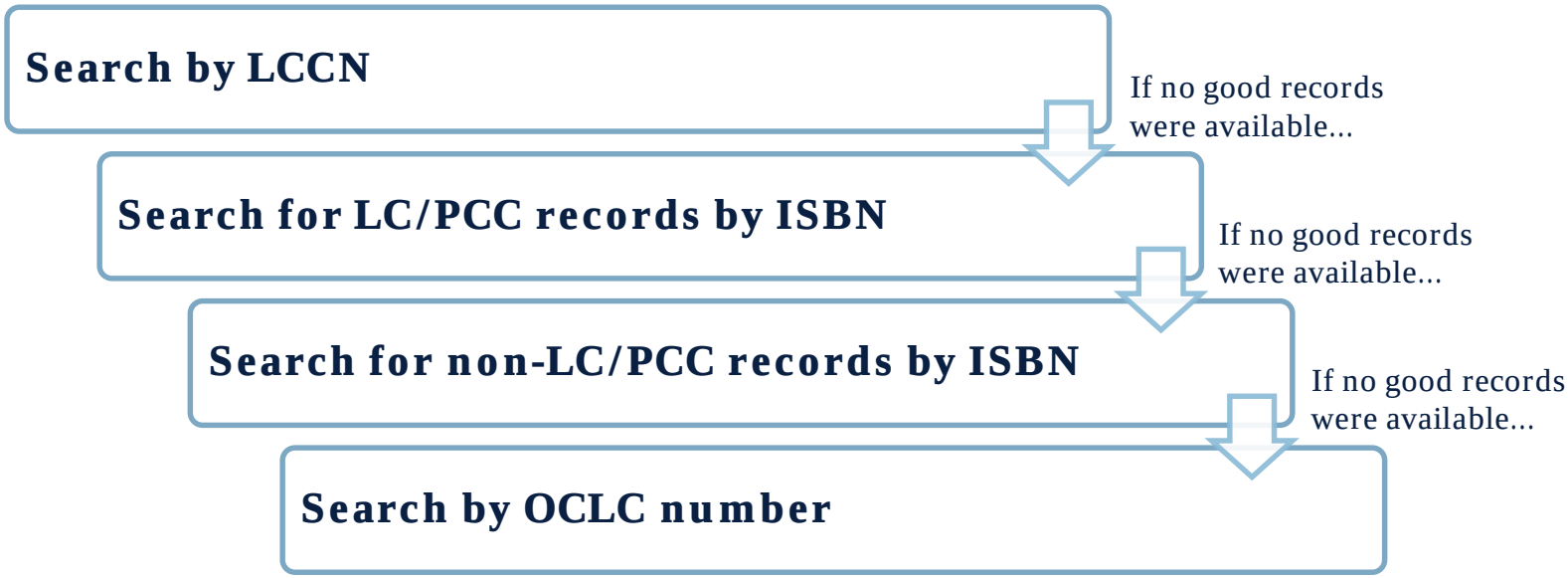
ISBN	NETWORK_NUMBER
1852865954 (pbk) :	(OCoLC)ocm33278093
9042015713	(OCoLC)ocm44770187
	(OCoLC)ocm03755470
086278588X	(OCoLC)ocm44603235
7810493590	(OCoLC)ocn793054015
	(OCoLC)71071850
7534411432	(CStRLIN)DCLP02-B328
7534411483	(CStRLIN)DCLP02-B331
2884643702 (pbk.)	(DLC) 2004430714
	(OCoLC)ocm00830370
	(OCoLC)ocm55125105

GREL
(General Refine Expression Language)

Regular Expressions

ISBN	NETWORK_NUM
1852865954	33278093
9042015713	44770187
	03755470
086278588X	44603235
7810493590	793054015
	71071850
7534411432	
7534411483	
2884643702	
	00830370
	55125105

Step 2 – Making API calls and parsing data



The process considers OCLC records with English cataloging language, encoding levels (blank, 1, 4, I, M), and LC subject headings (except for fictions) as good records

Output

BIB_ID	MFHD_ID	TITLE_BRIEF	BEGIN_	LANGUAGE	DISPLAY_CALL_NO	LCCN	ISBN	OCLC	SUBJECT	CALL_NO	ELvl
4038823	4405235	Greatest Superman stories ever told.	1995	eng	UNCAT008176		1852865954	60120126	Yes		M
4633817	5025559	Ecriture de la ruse	2000	fre	UNCAT127293	418550	9042015713	44770187	Yes	PQ645 .E37 2000	
4635677	5027449	Legend of Daniel Williams,	1958	eng	DIVUNCAT4635677			3755470	Yes	PZ3.F895 Le3	I
4688289	5081711	Power game : Fianna Fail since Lemas	2000	eng	UNCAT141182	455443	086278588X	44603235	Yes	JN1571.5.F5 C65 2000	
4732988	5127559	Xin Zhongguo mao yi si xiang shi	1999	chi	UNCAT182186	2001420544	7810493590	44810196	Yes	HF3836.5 .X56 1999	
6263085	6928559	Tshisekedi et le destin du Congo	2003	fre	UNCAT271409			71071850	Yes	DT658.2.T735	I
6350307	7007651	Lao Xianggang : dong fang zhi zhu	2000	chi	CHIUNCAT297393	2002325146	7534411432	47275624	Yes	DS796.H757 C46176 2000	
6350323	7007661	Lao Hangzhou : hu shan ren jian	2000	chi	CHIUNCAT297395	2002325148	7534411483	46716073	Yes	DS797.88.H364 L53 2000	
6478572	7118953	Saint Pierre Claver, apotre des negres	1893	fre	DIVUNCAT6478572			830370	Yes	PZ7 .M885Ha	I
6526807	7159007	Carte administrative de l'Algerie.	2003	fre	NEAUNCAT303272	2004335002		55125105	Yes	G8241.F7 2003 .C2	
6718623	7317035	Hong Fangzhou yan jiu lun ji	1998	chi	CHIUNCAT360891			43682117	Yes	DS753.6.H65 H65 1998	
6721631	7319517	Sathani Kaset Luang Ang Khang.	2000	tha	THAUNCAT947598			223071754	Yes		M
6785840	7366110	Cinema spiritualiste	2004	fre	UNCAT419393	2005397150	220407490X	57373072	Yes	PN1995.5 .A88 2004	4
7422557	7941297	Parochieregisters van O.L.V.-Waver	2000	dut	UNCAT801584			868153117	Yes	CD1688.5.A2 O67 2000	M
7857096	8309765	Chutplian prathet Thai	2006	tha	THAUNCAT947262			181141209	No (Fiction)		M
7909097	8353529	Zeng xiang quan tu San guo yan yi	1931	chi	CHIUNCAT557295			21655978	Yes		I
7953794	8387999	Botkhwam lang rang chut thi song	1971	tha	THAUNCAT94670			63991091	Yes		M
8090080	8504859	Tawq al-yamam : riwayah	2008	ara	NEAUNCAT562892	2008334756	9953682879	191758501	No (Fiction)	PJ7860.A224 T28 2008	
8153862	8559761	Lok thang bai hai nai khondieo	2001	tha	THAUNCAT946725			417243015	Yes		M
12898606	12990045	Sous les fleurs, des larmes	2015	fre	UNCAT966595	2016415083	9782350450568	953387584	No (Fiction)	PQ3989.3.K6485 S78 2015	
12943270	13032103	Sometimes : poetry	2015	eng	UNCAT894152	2015327637		940340936	No (Fiction)	PR9540.9.J37 S65 2015	
12948097	13036770	Instrumental queens	2015	eng	UNCAT894153	2015328126	9789696670001	927141484	No (Fiction)	PR9540.9.S59 I56 2015	I
12974492	13061551	Poets and their visions	2015	eng	UNCAT894170	2015363115	9789553065568	962412797	Yes	PR601 .W49 2015	
12979644	13065868	Sailing on the rock sea : puisi/poem	2016	eng	INDUNCAT934136	2016351906	9786027432833	972093049	No (Fiction)	PL5089.M74 B4 2016	I
13003035	13088238	Gengo no taisho kenkyu	1974	jpn	JPNUNCAT858467			33558108	Yes		M
13184283	13254060	Il pleut des avions : roman	2016	fre	UNCAT903334	2017363399		974567698	No (Fiction)	PQ3989.2.N28 P548 2016	
13201210	13267827	Shokugyo fujin ni kansuru chosa.	1924	jpn	JPNUNCAT861507			123357444	Yes		M
13203269	13269358	People's movements in Pakistan	2016	eng	UNCAT882521	2017336121		981761772	Yes	HN690.5.A8 K48 2017	
13206802	13272089	Digital customs progressive engageme	2016	eng	UNCAT894265	2015362445		946579483	Yes	HJ6609 .B3 2016	I
13208150	13273238	Making quality education a reality	2016	eng	UNCAT894272	2015515564		957656247	Yes	LA1162 .M35 2016	

Limitations

- Search exclusively based on LCCN, ISBN, and OCLC numbers
  → Some backlog titles still need manual searching
- Each API wskey has a limit of 50,000 queries a day
  → Be careful if spreadsheet has tens of thousands of rows
- Search does not check Dewey call number, FAST and other subject headings
  → Create your own operation by modeling ours
- Process is not perfect
  → Staff must verify the accuracy of the matches found

Yale University Library Fun Facts

- 15 million print and electronic volumes held
- 80K new accessions each year in MPS
- Over 12K students, 4462 international students from 118 countries
- Providing multi-language collections can be challenging for cataloging and discovery

Background – Cataloging Dilemma #1

- Technical Services moved out of the main library and off campus in Spring 2016.
    - *What to do with 33K cataloging backlog?*
- A third of the cataloging backlog was outsourced due to:
    - lack of language expertise
    - lack of room at new space
- 20K moved to new site and worked on as time permitted.

Keeping work home–Cataloging Dilemma #2

So…

How can we reduce future outsourcing?

Keeping cataloging at home:

- Lowers outsourcing costs for copy cataloging
- Outsourced materials require ongoing maintenance
    - records returned without call numbers
- Staff regain ownership of work
    - Less angst about outsourcing

Off to a good start

- Yukari's OpenRefine code used to run global report of all 20K+ backlog titles against OCLC to identify copy
- 50% of the titles searched were found to have copy
- This automation and rate of return helped to prove we could reduce the need for future outsourcing of cataloging backlog materials
- This week's report shows backlog is down to 14, 132 titles with 7048 with copy.

Yellow flags=copy

- Staff flag all titles found to have records in OCLC
- Students pull yellow tagged titles for staff processing as needed
- This eliminates unnecessary pulling of books with no copy

Southeast Asia Backlog Reduction Pilot

Purpose: to compare manual searching against OpenRefine use for eliminating Southeast Asia Collection backlog

- 3 staff members
    - 1 week of manual pulling of titles – 2 hours per staff, total 6 hours
    - 4 weeks of using OpenRefine – 5 hours per staff, total 60 hours

Collections in the Pilot:

- Indonesian
- Vietnamese
- Thai
- Burmese
- Malay
- Tagalog

Pilot Statistics

- 4446 titles identified in the backlog
- Manual Results: 200 titles searched, 65 cataloged , 6 hours
- OpenRefine Results = 713 cataloged, 60 hours

  Indonesian: 350 of 1510

  Vietnamese: 35 of 1141

  Thai: 79 of 630

  Burmese: 85 of 537

  Malay: 134 of 317

  Tagalog: 30 of 311

Success! SE Asia staff give a thumbs up to reducing their backlog

Roongtiwa Harlow, Elaine Pacelli, and Karen Van

Other uses for OpenRefine

- Eliminating record duplication for LC deposit titles we batch load from OCLC

- Eliminating searching for held titles from large selector wish lists

Want to try our OpenRefine code?

https://github.com/ysugiyama3/backlog_lookup

Thank you!

Yukari Sugiyama, Metadata Librarian

yukari.sugiyama@yale.edu

Debbie Falvey, Collections Procurement Librarian

debra.falvey@yale.edu